Broken promises

Intense public support for clinical research can be a mixed blessing — and the hunt for a vaccine against AIDS offers an important lesson for many biomedical initiatives on what can go wrong.

Last year's failed clinical trial for Merck's HIV vaccine led many to claim that AIDS vaccine research is in crisis. And at a meeting last week in Bethesda, Maryland, to examine how the field should proceed, discussion centred on the need for a "mid-course correction", as Warner Greene, director of the Gladstone Institute of Virology and Immunology in San Francisco, put it.

The large-scale trial for Merck's candidate vaccine was well-designed and had been launched with considerable optimism. Yet the vaccine not only failed to lower patients' viral loads, it actually boosted the risk of infection in some groups. And worse, there are few strikingly different vaccine candidates in the pipeline ready for human trials. The PAVE 100 trial for a vaccine candidate similar to Merck's has already been cut in size, and discussions are taking place on whether the trial should go ahead at all.

But if the AIDS vaccine field is in crisis, it is partly a crisis of its own making — one that holds lessons for many other areas of biomedical research. Decisions to move Merck's vaccine candidate and a previous failed candidate into clinical trials were based only partly on science. Also a factor was the field's need to show the public that progress is being made, thereby justifying the millions of dollars it receives from philanthropists and taxpayers. Historically, AIDS research has been afforded especially strong public support in the United States, but that means it works under an unusually intense public microscope. When promised results are not delivered, it's a double blow to the field, setting back the search for therapies while eroding public support.

This is an important lesson for other fields, most notably those, such as autism, Parkinson's and stem-cell research, that are funded by a new breed of philanthropic organizations and agencies that use a more 'business-like' management model. The end-point and milestone-driven funding model used by these organizations has become increasingly fashionable, as has their practice of setting more directed research priorities at the outset.

But there are consequences of the business-like approach that cannot be ignored. One is that trade-offs inevitably have to be made — often at the expense of basic research. The AIDS model provides a perfect example of this. When times are flush, as they were at the turn of this century, there is enough money to support clinical and basic researchers alike. When times become lean, as they are now, basic science loses out to the infrastructure that has been committed to clinical trials and 'big science' initiatives, such as the US$300-million Center for HIV–AIDS Vaccine Immunology. The US National Institute of Allergy and Infectious Diseases, the leading player in the US AIDS initiative, is now spending slightly more of its extramural AIDS-vaccine budget on testing vaccines than on discovering new candidates. As a result, the development pipeline has run dry of new ideas. To the agency's credit, it recognizes the problem and its director, Anthony Fauci, has pledged to "turn the knob in the direction of discovery" (see page 516).

> "Researchers do not serve themselves well by promising results they cannot deliver."

The other consequence of setting explicit milestones is the risk of public backlash when the results aren't delivered. AIDS vaccine research has made numerous promises it could not keep, such as when President Bill Clinton declared in 1997 that the nation would produce an AIDS vaccine within a decade. With every failure, public scrutiny grows. On 23 March, for instance, an article by the non-profit AIDS Healthcare Foundation appeared in *The Baltimore Sun* calling for an end to clinical trials for AIDS vaccines.

Yet rather than learning from AIDS research, other fields are repeating its mistakes. The California Institute for Regenerative Medicine, for example, has promised to deliver "clinical proof-of-principle that transplanted cells derived from pluripotent cells can be used to restore function for at least one disease" by the end of its ten-year mandate. If this goal isn't met, the state's public, which voted to create and fund the institute, is likely to wonder whether its investment was worthwhile.

Researchers have no choice but to reach out to the public for support. But when they do so, they do not serve themselves well by promising results they cannot deliver. ■

Double vision

The need to transform the world's energy technology is even greater than many perceive.

How much energy, and of what sort, should we expect the world to be generating in the decades to come? This is a question of crucial importance to economics, development and the management of climate change. But as the Commentary on page 531 argues, it is not an easy question to answer (see also News on page 508).

In 2000, the Intergovernmental Panel on Climate Change introduced a range of 'emissions scenarios' that sought to estimate the amount of carbon emitted over the twenty-first century for a range of different assumptions. Climate modellers have since used a small subset of these scenarios as the basis for their analyses of possible climate change. But these scenarios are not without their controversies. One is their much-criticized approach to currency exchange rates, which some economists say overstates the impact of industrialization in developing countries. Another, it turns out, is not so much a problem with the scenarios themselves, but with the way some of them are seen as being 'business-as-usual' — in other words, what would happen if we do nothing differently.

That perception leads to an apparently straightforward measure of

the climate-change problem for each scenario: take the anticipated carbon emissions and subtract an emissions profile designed to bring about a specific goal — say, a significant chance of keeping the increase in global average temperature at or below 2 °C. The result is the amount of carbon emissions that must be avoided to reach the goal. This figure, typically hundreds of billions of tonnes of carbon over the century, is then treated as the challenge that must be met by new energy technologies and efficiency drives.

But this approach brings with it a risk of double counting. The 'business-as-usual' scenarios actually assume that economies will somehow steadily increase their energy efficiency and decrease the amount of carbon they emit per unit of energy. The scenarios don't specify what technologies will be used to achieve these things, so unwary users could wrongly consider all reductions in either measure as progress over the status quo scenario. As a result, the amount of decarbonization required to achieve a stated goal is far larger than is often assumed.

The challenge is daunting enough without such misprision. An ever greater amount of world domestic product is coming from developing economies where the amount of energy used and carbon emitted per unit of economic activity is growing.

The history of the industrialized nations does suggest that there would, over time, be progress in energy technology and efficiency even without any threat of climate change. But global policy must now aim to increase this progress greatly, which means we have to accept that there are no meaningful 'business-as-usual' scenarios. The required progress requires new incentives, new research and, crucially, new costs imposed on the users of fossil fuels. It would also benefit from new economic models that, unlike the ones used with current scenarios, can show how the various policies aimed at replacing fossil fuels might work.

We also need to accept that the challenge of providing humanity with more energy in a minimally damaging way is even bigger than is usually thought. A simple calculation illustrates the point. If a growing population is to have access to the energy needed to drive development, then in 50 years the world's energy supply will need to at least double, even if energy use in tomorrow's great powers never rises to the per capita level seen in industrialized nations today, and if use in those nations is severely curtailed by efficiency improvements.

At the same time, if the world's climate is to be stabilized, carbon emissions from fossil fuels will have to be reduced dramatically in comparison with today's levels. To satisfy both those requirements, the amount of energy generated without the emission of fossil carbon needs to increase by something like a factor of ten.

That is the most daunting challenge humanity has ever sought to meet with a united front. The chances of success will be helped by models that allow different strategies to be compared. But no amount of scenario planning can replace the need for the will and leadership. ∎

Mismanaged measures

Surrogate end points can be helpful in clinical trials — but only if they are used with care.

"What gets measured, gets managed" is an adage that doctors know all too well. Physicians routinely monitor diseases by measuring biological factors such as cholesterol or blood pressure. And clinical trials are starting to use such indirect measures as 'surrogate markers' of a treatment's effect on the disease in clinical trials (see page 510). The practice has helped to get drugs to market quickly by allowing researchers to perform rapid and easy measurements rather than waiting for long-term clinical outcomes such as heart attacks or strokes.

But every so often, a surrogate deceives, as in the unexpected results from two recent clinical trials. One found that the drug ezetimibe, which lowers cholesterol levels, did not seem to slow the steady march of atherosclerosis in patients with high cholesterol (J. J. P. Kasterlein *et al. N. Engl. J. Med.* **358,** 1431–1443; 2008). In the other, one part of the trial was halted when the researchers found that using a combination of strict diet and insulin to lower diabetics' blood-sugar levels to those of healthy people actually increased mortality rates in some patients.

These results conjure up unpleasant memories of past surprises with surrogate markers. Several drugs that quelled irregular heartbeats in patients who had had a heart attack shocked cardiologists when they were found also to increase the risk of a subsequent heart attack. Another candidate drug for controlling cholesterol, torcetrapib, raised levels of the 'good' cholesterol — high-density lipoproteins — yet increased the number of deaths. Such events are troubling, and regulators face a high-stakes tug-of-war between the risk of relying on a surrogate marker, and the harm of delaying a potentially useful therapeutic from being used in the clinic.

Cardiologists at this week's annual meeting of the American College of Cardiology urged their colleagues to try other medications before resorting to ezetimibe. The warning should not have been necessary. Other cholesterol-lowering drugs have been on the market for 20 years, and have been vetted by large-scale clinical trials that evaluated direct outcomes such as heart attacks. Ezetimibe was approved five years ago in the United States solely on the basis of its ability to lower lipid concentrations. Large clinical trials to look at its effects on the cardiovascular system were not initiated until several years after it was approved. Some of the results aren't due until 2011. The choice between ezetimibe and the older drugs should have been clear, yet the statistics suggest that many doctors favoured ezetimibe. In 2006, ezetimibe accounted for 15% of the prescriptions for cholesterol medication in the United States. Sales of the drug climbed above $5 billion in 2007.

Why did the medical community embrace ezetimibe with such enthusiasm? There are two likely explanations. The first is an aggressive and effective marketing campaign. The second is a firm belief in the predictive power of a trusted surrogate. Previous cholesterol-lowering drugs had proven successful, and in heart disease as in diabetes, the belief came to be: 'the lower, the better'.

We won't know for some time whether ezetimibe works for the end points that matter: heart attack and death. But it is clearly time to check our assumptions. Surrogate end points are valuable, but are not always accurate indicators of the important outcomes. Drugs approved solely on the basis of surrogate end points come with a risk that should not be ignored, no matter how trusted the surrogate — or how persuasive the marketing campaign. ∎

RESEARCH HIGHLIGHTS

Surface surprise

Geophys. Res. Lett. **35,** L04605 (2008)

The efficiency with which photosynthetic phytoplankton removes carbon dioxide from the surface waters around Antarctica is itself dependent on the CO_2 concentration of those waters. This opens up the possibility that a negative feedback system exists, whereby increased plankton growth under high-CO_2 conditions results in proportionally more CO_2 being pumped out of surface waters and into sediments.

Philippe Tortell at the University of British Columbia in Canada and his co-workers found that diatoms — a common type of phytoplankton — normally seen in summer blooms in the Ross Sea were the greatest beneficiaries of this effect, with chain-forming diatoms doing particularly well. The finding may have implications for understanding the biogeochemistry of past ice ages as well as conditions to be expected in the future.

PANORAMIC/GETTY

MOLECULAR BIOLOGY
Hepatic hope

Nature Biotechnol. doi: 10.1038/nbt1396 (2008)

There are no approved therapies for the scarring that underlies cirrhosis of the liver. But delivering a small interfering RNA (siRNA) — a molecule that intercepts specific messenger RNAs — to the cells that produce the bulk of this collagen scar tissue could reverse the damage.

This approach worked in rats with liver damage when Yoshiro Niitsu and colleagues at the Sapporo Medical University School of Medicine in Hokkaido, Japan, administered an siRNA in fatty bubbles known as liposomes. These were coated with vitamin A, which tricked the cells into letting the drug in. The siRNA the authors used jams the expression of a crucial protein in collagen production.

A. WILLIAMS/ZEFA/CORBIS

NEUROBIOLOGY
Aftershock

Science **319,** 1842–1845 (2008)

When given small electric shocks, humans can learn to tell the difference between two mirror-image versions of the same molecule that previously seemed indistinguishable.

Wen Li and her colleagues at Northwestern University in Chicago, Illinois, asked people to sniff the right- and left-handed isomers of two substances. On sniffing a given isomer of one substance, the volunteers received a small electric shock. This 'conditioning' process altered the activity of the primary olfactory cortex, a region of the brain that processes smell. Afterwards, volunteers distinguished between the two versions of this substance better (50% of the time) than would be expected by chance (33%). The two versions of the second substance, used as a control, remained impossible to tell apart.

The work shows that humans have a more extensive smell capacity than was thought, but they don't normally need to use it.

IMMUNOLOGY
Antibiotic antibodies

J. Clin. Invest. doi:10.1172/JCI33998 (2008)

The discovery of functional antibodies against strains of *Salmonella* that do not cause typhoid raises hopes that a vaccine can be developed. In Africa, such strains kill up to 24% of infected children in communities in which appropriate antibiotics and blood-culture facilities are available.

Calman MacLennan of the University of Birmingham, UK, and his colleagues noted that *Salmonella* infection in Malawi is more common in toddlers under two years of age than in older children. The researchers report that healthy children over 16 months old produce antibodies that destroy non-typhoidal *Salmonella* as it divides in the blood; younger children do not.

Such strains are susceptible to other immune-system mechanisms when inside cells, but these do not stop the bacteria from proliferating in the blood. This further supports the potential usefulness of an antibody-stimulating vaccine.

NANOMECHANICS
Pulling teeth

Appl. Phys. Lett. **92,** 113103 (2008)

Imagine a tiny cog that spins without its teeth touching a driving rack. Theoretical physicists have worked out that two corrugated racks moving in the same direction on opposite sides of such a cog would cause the cog to turn backwards and forwards repeatedly. This set up might be used to make a low-wear mechanical oscillator, say MirFaez Miri of the Institute for Advanced Studies in Basic Sciences in Zanjan, Iran, and Ramin Golestanian of the University of Sheffield, UK.

The device's moving parts would need to be of nanoscale dimensions because the cog's motion relies on the Casimir force, a quantum-mechanical effect that pulls two surfaces together over very short distances. The Casimir force is highly sensitive to small changes in separation, so the device might also make a tiny position sensor.

INFORMATION SYSTEMS
Web spread

Proc. Natl Acad. Sci. USA **105,** 4633–4638 (2008)

The dissemination of a single piece of news over the Internet has been observed and analysed, revealing many more steps than theoretical models had predicted, and thus greater opportunity for information to be altered.

David Liben-Nowell at Carleton College in Northfield, Minnesota, and Jon Kleinberg at Cornell University in Ithaca, New York, studied variations of two e-mail petitions — one about opposition to the Iraq war and the other concerning government

funding of radio and television. Each recipient added their name before forwarding the petitions to friends. The variations therefore contained different but overlapping lists of names that revealed the paths by which they spread.

Instead of diffusing from person to person like an infectious disease, the two petitions moved around the Internet in similar and surprisingly narrow tree-like patterns, and continued to reach new people even after several hundred steps.

CONSERVATION BIOLOGY
Embattled bats

J. Mammal. **89,** 18–24 (2008)
Six colonies of Brazilian free-tailed bats (*Tadarida brasiliensis*) in the southwestern United States may have plummeted from 54 million members to 4 million since 1957, according Margrit Betke at Boston University in Massachusetts and her colleagues. This is bad news for humans, because the bats eat the moth *Helicoverpa Zea*, the larva of which is one of North America's most destructive farm pests.

The team used heat-sensitive video cameras to pick out the creatures' warm thoraxes in the dark to record colonies emerging en masse from caves. Because fast-moving bats occlude one another only briefly, comparing sequential video frames allowed the authors to produce a reliable census.

The 1957 study was much less accurate as it relied on human observers, but massive population decline due to guano mining and heavy DDT use in the 1960s seems likely.

PHYSICS
Record breaker

Nature Phys. doi:10.1038/nphys919 (2008)
The record for the most bits of information transmitted in a single flicker of light has been beaten by researchers at the University of Illinois at Urbana–Champaign.

Light can transmit quantum information twice as fast as zeros and ones can move through classical electronic devices. This is due to a process called entanglement, whereby the spins of a pair of photons are manipulated. But the capabilities of light-measuring instruments have limited rates to 1.585 bits per photon.

Julio Barreiro and his colleagues achieved an average rate of 1.63 bits per photon by further entangling the photon pairs, using a corkscrew motion of light known as its orbital angular momentum. Perfect measurement of this method could result in rates as high as 2.81, they say.

ECOLOGY
Carbon conundrum

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0800568105 (2008)
As atmospheric carbon dioxide concentrations rise, soya-bean plants defend themselves less well against some beetles, find May Berenbaum and her colleagues at the University of Illinois at Urbana–Champaign.

The researchers studied Japanese beetles (*Popillia japonica*) and soya-bean plants (*Glycine max*) in open-air laboratories in which the percentage of CO_2 in the air can be controlled. Crops grown in CO_2 concentrations of 550 parts per million — roughly twice the pre-industrial level — had reduced expression of genes involved in defensive signalling. Faced with lower levels of compounds that normally deter them, the beetles ate more soya-bean leaves.
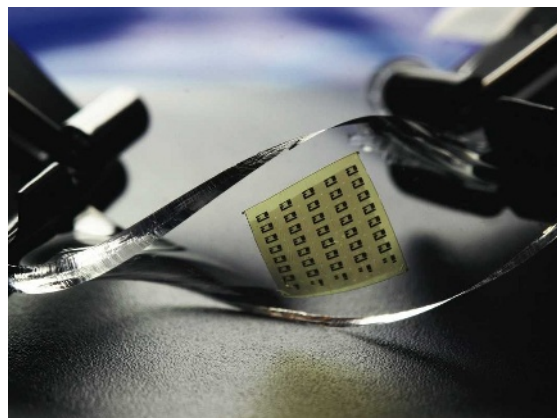
SCIENCE

MATERIALS SCIENCE
Bendy electronics

Science doi:10.1126/science.1154367 (2008)
Stretch them, fold them, unfold them, and flexible new integrated circuits (pictured above) still perform as well as rigid ones. They can be used to build logic gates, amplifiers and transistors.

The devices are made from arrays of unbroken silicon crystal nanoribbons, through which electrons flow rapidly. These are printed on an ultrathin layer of a polymer, on a substrate covered with 'sacrificial' glue that is later dissolved. The product is then transferred onto a stretched sheet of poly(dimethylsiloxane), another polymer, which makes the circuits stretchable when the strain is released.

Authors John Rogers of the University of Illinois at Urbana-Champaign, Yonggang Huang of Northwestern University in Evanston, Illinois, and their colleagues propose using the bendy circuits in wearable health monitoring equipment.

JOURNAL CLUB

A. P. de Silva
Queen's University, Belfast, Northern Ireland

A chemist looks at DNA-based molecular logic.

Logic gates — AND, OR and NOT gates — are used in all manner of electronic devices, for example computers, in which they are connected in huge arrays. Several research groups, including my own, have designed and built molecular logic gates since the early 1990s. But the usefulness of our efforts has been limited because linking these gates in series has proved difficult.

Recently, Reza Ghadiri and his colleagues at the Scripps Research Institute in La Jolla, California, constructed a full set of logic gates that release a single-stranded DNA sequence when provided with the correct combination of single-stranded DNA inputs (B. M. Frezza *et al. J. Am. Chem. Soc.* **129,** 14875–14879; 2007). This means that the output of one gate can be the input for another, and that the gates can be 'wired together' into multi-level circuits using the solution containing the DNA as a communication medium.

The gates work as follows. When a single strand of DNA pairs with a longer strand, an 'overhang' of unpaired DNA is left. If another complementary strand then comes along that is the same length as the longer strand, the overhang provides a foothold, allowing the new strand to push the shorter one off and form a full-length hybridized pair. Ghadiri *et al.* attached the DNA to beads so that different gates could be kept apart until the correct input was ready. They also added a fluorescent part to the final output signal to make the result easy to monitor.

This may not seem like much of an achievement to a computer buff. Nevertheless, I think the principle that this paper describes could pave the way to more useful molecular logic gates. In the meantime, the simple molecular logic gates that are available can serve in real-life applications such as identification tags for small micrometric objects. Semiconductor identification devices are too big for this purpose.

Discuss this paper at **http://blogs. nature.com/nature/journalclub**

507

NEWS

Are the IPCC scenarios 'unachievable'?

The Intergovernmental Panel on Climate Change (IPCC) has grossly underestimated the challenges of reducing and stabilizing greenhouse-gas emissions, according to an influential group of climate-policy experts.

The scenarios produced by the IPCC assume that very substantial technological advances — leading to greater energy efficiency and reduced carbon dioxide intensity — will happen spontaneously, without extra policy measures (see page 531).

Roger Pielke, Tom Wigley and Christopher Green argue that this is a "dangerously optimistic" assumption. To show its effects, the trio offer a contrasting 'frozen-technology' scenario, which assumes that future energy needs are met with technology available at the baseline year. They say that this demonstrates a need for new energy technologies as much as four times greater than that which seems to be required looking at some IPCC scenarios. *Nature* gets some reactions.

C. CALVIN; MCGILL UNIV.

S. KHAN/AFP/GETTY IMAGES

Bert Metz, co-chair of IPCC Working Group III
The claim that the IPCC has underestimated the technological challenge of stabilization is unwarranted and must be rejected. The fact that technological change is already significantly included in the reference scenarios is clearly stated in our report and its Summary for Policymakers.

The embedded technological change in the reference scenarios included in the IPCC's Special Report on Emission Scenarios was based on historical information. The assumptions about the rate of technological change in these scenarios have been thoroughly reviewed and are accepted by the community of technological-change experts. They confirm well-known facts about, for instance, the enormous improvements in computers over much shorter time-frames than expected. The assumptions also reflect that high economic growth normally goes hand in hand with high rates of technological change.

Fears that actual economic growth and energy use may develop in different ways than assumed in our scenarios are more justified. If so, it could be more difficult to reach low stabilization levels. The IPCC in 2006 initiated the development of new long-term scenarios. New scenarios will be developed over the coming years.
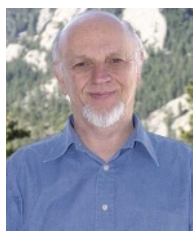
Richard Tol, energy economist, Economic and Social Research Institute, Dublin
This is a very valuable piece of education. Indisputably, in the IPCC's reference scenarios there is a lot of spontaneous technological progress going on already. Experts know this, but many people who write about climate change, or have to make policy decisions, don't.

Improving energy efficiency is not enough. Engineers could do it, but would need to work at least four times as hard to stabilize emissions.

The IPCC scenarios developed in 2000 don't match historical observations. For instance, nobody foresaw the rapid growth in recent years of China's economy, or the launch of the US$2,500 people's car in India. There has also been less technological change in past decades than the scenarios would have us believe will happen in the future, and hence a downward bias in the cost of emission reductions. We need new scenarios — not just any, but scenarios that are in line with the real development of global energy systems.

John Reilly, Massachusetts Institute of Technology Joint Program on the Science and Policy of Global Change, Cambridge
No question, these reference

Carbon-trading market has uncertain future

The budding carbon-offsets market could already be on its last legs, industry representatives say, if the latest European proposals are agreed. European negotiators went into a United Nations climate meeting in Bangkok this week warning developing countries that they need to step up to the challenge of climate change if they are to see additional money flowing into clean-development projects.

The idea that developing nations should be able to increase emissions for a time to grow their economies and lift their citizens out of poverty is grounded in the current climate treaty. But there is increasing recognition that industrialized countries — responsible for most of the current greenhouse-gas emissions — cannot alone address climate issues, given the rapid rise of emissions in emerging economies.

The European Commission (EC) weighed in on the issue in January with a set of policy proposals that would scale back the clean development mechanism (CDM), a programme under the Kyoto Protocol that allows businesses in rich countries to offset their emissions by funding clean-energy projects in developing countries. The proposals would essentially cap the CDM at current levels until 2020 if a new climate treaty is not agreed. Commission officials say that a moderate expansion would be allowed if an agreement is reached, although estimates vary as to the actual impact.

The recommendations are intended to boost emissions reductions at home in Europe, and at the same time give developing nations incentives to do more than sit back and watch the money flow in. Business interests say that the commission's proposals have thrown the entire carbon-offsets industry — valued at around US$18 billion last year — into doubt. "The cap, as it is designed now, will not provide any incentive for people to design new projects," says Michela Beltracchi, European policy coordinator for the International Emissions Trading Association, based in Geneva, Switzerland, which represents a range of business interests. "Effectively, the market will be killed."

How to help developing nations chart a cleaner path to industrialization is central to this week's climate talks, which mark the beginning of a two-year process of negotiating a follow-up agreement to the Kyoto Protocol, a course set at a previous meeting in Bali, Indonesia, in December 2007. Advocates of the CDM face a steady flow of criticism that the programme is not reducing emissions but merely transferring wealth to the developing world and, primarily, to China. Many of the emissions-reduction projects in the developing world that are registered as CDM projects would have gone ahead anyway, critics point out.

SHOCKS MAKE PLANTS PRODUCE MORE CHEMICALS
Reaction to stress leads to larger chemical yields.
www.nature.com/news

scenarios are dated and underestimate the amount of warming that will occur without new climate policies.

We at MIT have argued for probabalistic projections, but the IPCC scenarios followed a storyline approach that they explicitly stated should not be interpreted in terms of likelihood. This subtlety is often lost when the scenarios are rolled out.

Clearly, emissions have grown much more rapidly than projected. In our business-as-usual projection, technological advances that improve efficiency are more than offset by growing demand, especially in developing nations. An accurate picture requires careful assessment of various economies rather than a broad-brush numerical exercise at the global level.

Detlef van Vuuren, emission scenario developer, Netherlands Environmental Assessment Agency, Bilthoven
Emission scenarios are built around historical observations. Energy efficiency has in the past improved without climate policy, and the same is very likely to happen in the future. Including unprompted technological change in the baseline is thus logical. It is not very helpful to discredit emission scenarios on the sole basis of their being at odds with the most recent economic trends in China. Chinese statistics are not always reliable. Moreover, the period in question is too short to signify a global trend-break. Rapid economic growth, in combination with the high price of oil, might spur long-term developments in renewable-energy technologies, for instance.

Marty Hoffert, former chair of the Department of Applied Science, New York University
This analysis is long overdue. We're under a delusion that we will solve the problem of climate change casually. But what we have — cap-and-trade systems and the like — is plainly insufficient. We need a massive engineering effort, the size of the Manhattan Project.

Ottmar Edenhofer, IPCC Fourth Assessment Report lead author, Potsdam Institute of Climate Impact research, Germany
The dramatic conclusions grossly overshoot the mark. It is wise to ask whether the IPCC's assumptions on spontaneous versus policy-induced technological change are sensible. But freezing technological change is a mere thought experiment. Given the past, a frozen-technology baseline is extremely unlikely.

Global energy efficiency and intensity have improved over the past 30 years. China in the first half of the decade is a special case owing to structural economic changes, although it is indeed worrying that neither carbon nor energy intensity in China has decreased since 2001.

It is a misapprehension that the costs of mitigating climate change will be much higher than assumed. Our studies show that the transformation towards a low-carbon energy system is possible if appropriate policy instruments are implemented. As the price of fossil energy increases, energy-efficiency gains are likely to materialize, even without extra policy measures.

Robert Socolow, Carbon Mitigation Initiative, Princeton University, New Jersey
The case for aggressive action to stabilize atmospheric CO_2 concentrations at less than double pre-industrial values is far more robust than this commentary suggests. Politicians rightly associate such targets with deep cuts in annual CO_2 emissions in industrialized countries relative to today. The commentary seems to argue that the case for aggressive action is not strong unless the world in the absence of attention to carbon would emit far more CO_2 than the IPCC predicts.

The authors demonstrate remarkable self-confidence in dismissing the IPCC's consensus estimates of future carbon intensity. The premise of the IPCC's econometric models is that the past is a guide to the future. If I were constructing a revisionist case for very high future emissions, I would instead argue that the IPCC underestimates world economic growth.

Current attempts to mitigate climate change are based on evolutionary but unprecedented expansion of tested strategies, including renewables and nuclear power, CO_2 capture and storage, methane emissions reduction, and forest protection. No one can be certain these efforts will be sufficient. The revolutionary should be pursued in parallel with the evolutionary. It is foolhardy to undermine indispensable commitments to evolutionary measures.

To the authors, the IPCC report is a lullaby. To most of us, it's a bugle call. ∎

Quirin Schiermeier with additional reporting by Jeff Tollefson
See Editorial, page 503.

"There are voices saying that maybe CDM is not as credible as it should be, and we need to take those concerns seriously," says Artur Runge-Metzger, who oversees climate issues at the EC. He says that Europe needs to spur new technologies now because simply paying for offsets elsewhere won't solve the problem.

Miles Austin, head of European regulatory affairs for carbon-offsets dealer EcoSecurities, Dublin, is sceptical that a threat to pull the plug on the CDM will have much of an effect on the climate talks. "The CDM could fold and China wouldn't even notice," Austin says. He points out that all the projects in China — home to roughly half of the CDM credits — represent less than 1% of the country's annual growth.

Business interests are working to preserve and expand the CDM programme as the commission proposal moves through the European Parliament this year. Austin is promoting one amendment that would leave the overall cap on European emissions in place while carving out a new space for international credits. He

says that the proposal would ensure that investments in new technologies are made at home while allowing the CDM market to expand.

The big question for the future is how international credits will be handled by the United States, where lobbying is under way to shape a market that could dwarf that of the European Union. The leading climate legislation in the Senate would limit international offsets to 15% of the overall market, but as the legislation is currently written, those credits would need to come from another market, such as the

European trading scheme. This has led critics to suggest that the US legislation would essentially allow other nations to 'launder' international credits with no oversight from the United States, thereby preventing the country from influencing an important international debate.

Mike Wara, who studies emissions markets at Stanford University in California, says that the CDM programme is in need of reform as a minimum, and might not be the right model to solve many problems. For instance, developers are currently pushing for CDM credits for hydroelectricity, wind power and, most recently, natural-gas-fired power plants in China, Wara says, but all these are part of a broader national strategy. "I think the jury is still out about whether CDM is the way to go," Wara says. "It starts to look like the CDM is kind of like a production subsidy. It's great for feedlots and it's great for landfills, but do we really want the entire Chinese energy sector in the CDM? Probably not." ∎

Jeff Tollefson

PUNCHSTOCK

B. SACHA/CORBIS

SPECIAL REPORT

D. GRUBBS/AP

Drug markers questioned

A recent spate of worrying clinical-trial data has researchers questioning drugs approved on the basis of how they affect biomarkers rather than clinical endpoints. **Heidi Ledford** looks at surrogate markers.

Forty years ago, a debate raged among cardiologists over the nature of the link between high blood pressure and heart disease. Some believed high blood pressure was a causal factor in producing heart attacks and strokes. But others argued that the body raised blood pressure as a way of mitigating poor vascular health. The stakes were high: if increased blood pressure was a coping mechanism, then drugs that lowered it could be harmful.

After dozens of clinical trials, millions of people now take drugs to treat their hypertension. High blood pressure has become an accepted surrogate marker for cardiovascular disease, meaning that a fall in blood pressure can be used as an endpoint in clinical trials for drug approval. An advantage is that blood pressure can be measured quickly and cheaply, so a drug can be approved for use without the need to wait for its effect on distant and infrequent clinical outcomes such as heart attacks.

Drugs that lower blood pressure are now regularly approved for cardiovascular disease on the basis of their ability to fight hypertension alone — in the United States and Europe, for example, regulatory agencies do not require additional clinical trials to determine whether such drugs also reduce heart attacks or strokes.

Drugs for other conditions, from cancer to diabetes, are also often approved on the basis of such surrogate outcomes as reduced tumour size and a drop in blood-sugar levels.

But a recent spate of disquieting clinical trials involving high-profile drugs such as Avandia (rosiglitazone) is prompting researchers to re-evaluate this reliance on surrogate markers (see 'Surrogates under suspicion'). "It's been a watershed year," says Harlan Krumholtz, a cardiologist at Yale University. "It's shaking assumptions about how we should be evaluating these drugs."

Nobody is saying that surrogate indicators should be abandoned altogether. Many well-known drugs on the market today, including the breast cancer drug Gleevac (imatinib) and HIV-fighting antiretrovirals, were approved based on surrogate endpoints, and scandals are rare. "You don't want to deprive the public of a drug that possibly works," says Arthur Schatzkin, chief of nutritional epidemiology at the National Cancer Institute in Bethesda, Maryland. "On the other hand, you don't want to put something on the market that is harmful. This is the kind of thing that people at the FDA [Food and Drug Administration] tear their hair out over."

One difficulty is the struggle to show that a surrogate is causally related to disease outcome, rather than merely a correlate. Just because patients with flu have a fever, for example, doesn't mean that treating the fever will clear the infection. Absolute confidence in a surrogate requires complete knowledge of the pathways leading to disease. "The surrogate outcome would have to be involved in all pathways that might link the treatment to the clinical outcome," says Ross Prentice, a biostatistician at the Fred Hutchinson Cancer Research Center in Seattle, Washington. "Really we aren't ever in a position to know whether those conditions are strictly satisfied." That is particularly true of complex conditions, including cardiovascular disease and diabetes.

Regulatory bodies such as the FDA in some

Surrogates under suspicion

Avandia (rosiglitazone) was one of the biggest names to rock the pharmaceutical industry last year. The diabetes drug — GlaxoSmithKline's second-biggest seller — had been fast-tracked through the approval system on the basis of a surrogate marker of lowering blood-sugar levels. In May, a meta-analysis showed that Avandia increased the risk of heart attacks[1]. Several other drugs in the same class of glitazone compounds have yielded similar results.

This week, at a meeting of the American College of Cardiology in Chicago, Illinois, where clinicians gathered to discuss surrogate markers, there was more bad news. John Kastelein,

of the Academic Medical Center in Amsterdam, and colleagues presented clinical trial results for the cholesterol-lowering drug Vytorin (ezetimibe/simvastatin), produced by Merck and Schering-Plough[2].

Blood cholesterol is used as a surrogate for risk of cardiovascular disease and stroke. Patients taking Vytorin had reduced low-density lipids — the so-called 'bad' cholesterol — in their blood. But they showed no significant difference in the thickness of their carotid-artery walls compared with patients taking older cholesterol-lowering

drugs that had undergone more rigorous safety trials. Thickening is thought to be a surrogate for the risk of heart attack and stroke because it reflects a build-up of artery-blocking atherosclerotic plaques. There are no large-scale data on the frequency of heart attacks or strokes for Vytorin — those results are expected from another trial in 2011.

There have been several such results involving drugs approved on the basis of surrogate outcomes, with more expected, but none of the findings has overthrown dogma. The Vytorin results, for example,

did little more than pit a surrogate marker (lower blood lipids) against another surrogate marker (thickness of an arterial wall).

Prediman Shah, a cardiologist at the Cedars-Sinai Heart Institute in Los Angeles, California, questions the utility of artery-wall measurements, noting that the composition of a thickened wall differs from that of an atherosclerotic plaque.

Many researchers acknowledge that lowering cholesterol levels can't always protect against cardiovascular disease. A large-scale women's study showed that post-menopausal hormone therapy did an excellent job of lowering cholesterol but held no benefit

NEWSCOM

Cancer drugs approved on the basis of a surrogate marker of tumour size don't always reduce mortality.

cases allow drugmakers to bypass long-term safety studies, and conditionally approve a drug based on small studies using a surrogate marker, on the basis that future studies will be done to establish its clinical efficacy and safety. The fear here is that unanticipated side effects are missed. "Wait long enough, and you're going to find that all surrogates eventually fail due to these off-target effects," says Steven Nissen, a cardiologist at the Cleveland Clinic in Ohio.

And the FDA has a poor history of enforcing post-marketing trials — a 2007 report noted that drug companies had started only 29% of the post-marketing safety studies they had agreed to as part of a drug approval decision. By

dragging their feet on the required clinical trials, drug companies can make billions of dollars on a flawed drug. "If the drug withstands the test of time, then that's wonderful," says Maha Hussain, an oncologist at the University of Michigan Health System in Ann Arbor. "But you have to actually collect the data on survival."

It is unclear whether the current scrutiny of surrogate endpoints will affect how drug-approval agencies view them; the field has weathered scandals in the past. For now, at least, the FDA has not announced any plans to change its policies. In February, Senator Charles Grassley (Republican, Iowa) requested an investigation into how the FDA uses surrogate endpoints in drug approvals.

The issues are likely to become more pressing as projects to collect biomarkers of disease bolster the collection of potential surrogates. There is intense interest in finding surrogates for clinical trials of drugs for diseases such as cancer or Alzheimer's. "If you're trying to prophylax against some dreadful disease that won't show up in 20 years," says Robert Temple, director of the FDA's Office of Medical Policy in Rockville, Maryland, "you almost don't have a choice." ■

1. Nissen, S. E. & Wolski, K. *N. Engl. J. Med.* **356,** 2457–2471 (2007).
2. Kastelein, J. J. P. *N. Engl. J. Med.* **358,** 1431–1443 (2008).
3. Prentice, R. L. *et al. Am. J. Epidemiol.* **163,** 589–599 (2006).
4. Cardiac Arrhythmia Suppression Trial Investigators *N. Engl. J. Med.* **321,** 406–412 (1989).
5. Cardiac Arrhythmia Suppression Trial II Investigators *N. Engl. J. Med.* **327,** 227–233 (1992).
6. Temple, R. J. in *Clinical Measurement in Drug Evaluation* (eds Nimmo, W. S. & Tucker, G. T.) (Wiley, New York, 1995).

for heart disease risk[3].

Other surrogate markers also look shaky. In February, researchers halted a clinical trial when they found that in patients with diabetes, those undergoing a severe regime of drug treatment and dietary restriction to drive down their blood-sugar levels were at greater risk of dying than those whose blood sugar was kept at slightly higher levels. The notion that, for diabetics, the lower the blood sugar, the more effective the therapy, is now being questioned.

"The history of clinical trials tells us that you can be burned badly if you use surrogate endpoints that have not been validated against clinical endpoints," says Shah. In the 1980s, the Cardiac Arrhythmia Suppression Trial tested three antiarrhythmia drugs in patients who had an irregular heartbeat after a heart attack. At the time, this was so strongly accepted a treatment that some questioned whether it was ethical to withhold the drugs from the placebo group.

Instead, the drugs more than doubled the risk of death or heart attack[4,5]. "Everyone was so confident that if you quieted the extra heartbeats, the patients would do better," says Harlan Krumholtz of Yale University. "But people died." It is now thought that the drugs, called sodium-channel blockers, might worsen irregular heartbeats when blood flow to the heart is limited, as could be the case in patients with a history of heart disease. Three years later, Robert Temple, director of the Office of Medical Policy at the Food and Drug Administration in Rockville, Maryland, wrote[6]: "It is not easy to think of a greater medical error, since the practice of therapeutic bleeding, than the use of antiarrhythmic drugs in patients after myocardial infarction."　　H.L.

EPA feels heat over flame retardant

A much-anticipated report on the health hazards of a ubiquitous flame retardant has been delayed amid controversy over the removal of a respected toxicologist from the US Environmental Protection Agency (EPA) advisory panel reviewing the report.

The report, which was due last week, is expected to mount pressure on the chemical industry to ban decabrominated diphenyl ether (deca-BDE), which is used as a flame retardant in furniture, carpets, and televisions and other electronic goods. The EPA has delayed its release by a month and experts tracking the issue expect further delays. "It's been stuck at the White House Office of Management and Budget for months," says a frustrated EPA scientist who asked not to be named.

In February it emerged that the EPA had axed one of its senior toxicologists, Deborah Rice, from the panel's chair after an industry group, the American Chemistry Council (ACC), complained about her. A Freedom of Information Act request revealed that Rice was dismissed last year after the ACC questioned her objectivity. According to the EPA source, George Gray, an EPA official, removed Rice and struck all of her comments from the draft report.

Rice, now at the Maine Center for Disease Control and Prevention (CDC), had published research showing that deca-BDE can affect brain development in animals (D. C. Rice *et al. Neurotoxicol. Teratol.* **29,** 511–520; 2007). As a Maine CDC employee, Rice testified about these harmful effects before the Maine state legislature, which subsequently banned the chemical (see 'Banned compounds'). Washington state had already banned deca-BDE, and other states are considering it. A 1 April court ruling has forced the European Commission to ban deca-BDE. Denmark and Sweden have also banned it.

Deca-BDE is a flame retardant used in common household goods from televisions to carpets.

Rice's dismissal has incensed toxicologists within and outside the EPA. "In a word, it's outrageous," says Thomas Webster, who studies deca-BDE at the Boston University School of Public Health in Massachusetts. "She is a top toxicologist and her removal from the panel is utterly ridiculous."

One high-level EPA scientist who requested anonymity told *Nature*: "Many of us who work here are really angry about this."

Politicians are now stepping into the fray. On 17 March, the governor of Maine, John Baldacci, asked the EPA to reinstate Rice as chair of the panel and return her comments to the draft report. That same day, two Democratic Congressmen from Michigan, Bart Stupak and John Dingell, demanded an explanation from EPA administrator Stephen Johnson. "The ACC does not assert that Dr Rice had any pecuniary interest in the human health assessment at issue, and therefore seems to argue that

scientific expertise with regard to a particular chemical ... is a basis for disqualification from a peer review board. This does not seem sensible on its face," the Congressmen write.

Their letter also names nine other members of EPA external review panels who either work for or receive funding from manufacturers of chemicals those same panel members were reviewing. "The routine use of chemical industry employees and representatives in EPA's scientific review process, together with EPA's dismissal of Dr Rice, raises serious questions" about how the EPA applies its conflict-of-interest rules, the letter states.

The letter asks for a response by 27 March, but on that day EPA associate administrator Christopher Bliley replied that "because of the large number of questions and requests for documents ... we are not able to provide you with a response at this time". ■

Brian Vastag

Banned compounds

Concern over the polybrominated diphenyl ether (PBDE) class of compounds surfaced in 1998, when Swedish researchers reported finding the compounds in stored breast milk. From 1972 to 1997, the concentrations of these chemicals in human milk had shot up 60-fold.

Toxicologists feared that the compounds' structural similarities to toxic, banned polychlorinated biphenyls (PCBs) meant the flame retardants might have a similar

effect on the body.

Other work found that Americans carry much higher PBDE concentrations in blood and milk — as much as 100-fold higher, in some studies — than Europeans. In addition, environmental sampling shows that the compounds persist for decades, finding their way into birds of prey and even polar bears.

In 2004, in the face of growing toxicity data, the chemical industry agreed to stop making
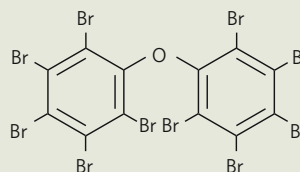
Deca-BDE (left) is similar to toxic, banned polychlorinated biphenyls.

two classes of PBDEs, penta-BDE and octa-BDE, so named for the number of bromine atoms on their rings. But manufacturers continue producing large quantities of deca-BDE, the fully brominated version of the basic molecule.

Manufacturers claim that
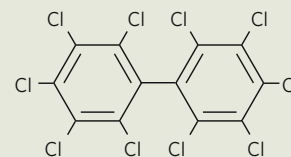
deca-BDE saves 300 lives a year in the United States by preventing or slowing house fires. But some researchers dispute that figure. "It's really difficult to know how many lives are saved by it," says Daniele Staskal, a toxicologist at the consulting firm ChemRisk. **B. V.**

NUMBER CRUNCH

1979 is the year in which
the World Health Organization
declared smallpox officially extinct.

87% of Scottish participants in
a recent public survey conducted
by the Society for General
Microbiology in Reading, UK,
had never heard of this medical
triumph.

40% of 16–24-year-olds in the
survey did not even know what
microbes are.

ON THE RECORD

**"I know a lot about
sheep and cattle;
I don't know much
about satellites. But
I would say it is a fuel
cell off some stage of a
rocket."**

James Stirton,
whose farm in
Australia was hit by
some space debris
last year, clearly
knows more
about space
aeronautics than
he thinks.

Sources: *The
Register, Wired,
Society for General
Microbiology,
Reuters*

REUTERS

SIDELINES

Easy ways to other Earths

A more precise way of calibrating the measurement of spectra should make it possible to identify Earth-sized planets around stars outside our Solar System using ground-based telescopes.

The technique, described on page 610 of this issue, makes use of laser 'combs'. Such combs are fine-toothed spectra of light that, when used in synch with atomic clocks, give scientists an exquisitely defined and stable reference point for measuring the wavelengths of light. The inventors of the laser comb shared half of the 2005 Nobel Prize in Physics, and the technology has made its mark in areas as diverse as chemical sensing and telecommunications. Now it can add astronomy to its list of applications.

'Astro-combs' should allow astronomers to measure the spectral lines of starlight with a precision as much as 60 times greater than the current state-of-the-art technique, say the authors of the paper, who are based at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts. As planets swing around stars, they induce movement that causes slight shifts in the stars' spectral lines. Astronomers have been using those shifts to infer the existence of planets for more than a decade, but until now have only been able to pick up the relatively large shifts caused by Jupiter-sized planets. The astro-combs should allow velocity measurements as precise as 1 centimetre per second, which brings the smaller shifts caused by Earth-sized planets into the detectible range.

Extending this technique's sensitivity will improve astronomers' picture of what other stellar systems look like. It will also provide targets for future work aimed at characterizing the atmospheres of any Earth-sized planets discovered, and at looking for indications of habitability.

Planet hunting is not the only task that the combs will make easier. If fitted with comb-based systems, the coming generation of giant telescopes now on drawing boards might, over a period of decades, make direct measurements of the cosmic acceleration put down to 'dark energy'. "The impact of the technology will be huge," says Ronald Walsworth, a physicist at the Harvard-Smithsonian Center and a co-author of the astro-comb paper.

By June, the Harvard-Smithsonian group should have started testing its prototype system at the Multi-Mirror Telescope (MMT) Observatory on Mount Hopkins in Arizona. And, in 2009, the researchers plan to set up a planet-spotting system at the 4.2-metre William Herschel Telescope on La Palma, in the Canary Islands, in collaboration with the Geneva Observatory.

A rival group of researchers from the European Southern Observatory (ESO), which is based in Garching, Germany, says they have already used a similar astro-comb system to gather data with the Vacuum Tower Telescope on the neighbouring Canary Island of Tenerife. "Already we're showing that using these things on an astronomical telescope is a reality," says Michael Murphy, an astronomer at Swinburne University of Technology in Melbourne, Australia, and a collaborator with the ESO group. The ESO team plans to use an astro-comb system at the observatory's 3.6-metre telescope at La Silla in Chile before eventually installing them at its four 8.2-metre telescopes at Cerro Paranal, also in Chile. "We came in as the fast, quick Americans who figured out a way to do just-good-enough calibration right away," says Walsworth of the Harvard team. "They're taking the Mercedes-Benz approach."

The Herschel telescope might find alien Earths.

D. PARKER/SPL

Whatever the approach, it will be a challenging technique to apply. For a start, planets are not the only things that move the surfaces of stars; they roil with starspots and pulse with starquakes. Such noise could render astro-combs moot, worries David Latham, a Harvard-Smithsonian astronomer. "It is a very large leap of faith that we can correct out the jitters in the stars themselves," he says. The combs' precision also makes strenuous demands on the stability of the spectrographs themselves. And because more photons are needed, more precise measurements will require either longer observation times or bigger telescopes.

Velocity measurements are not the only way to detect planets; they can also be seen by the changes they induce in the stars' position in the sky, and, in some cases, by the dimming they induce as they pass between their star and Earth. NASA's Kepler mission, due to launch in early 2009, aims to monitor the brightness of

STORIES BY SUBJECT
Find all our space and
physics news online
**www.nature.com/news/
archive/subject/index.
html**

about 100,000 stars in a particular patch of sky in the hope of measuring such transits, something France's Corot satellite is already doing on a smaller scale. However, the stars Kepler will be looking at are almost all too distant for follow-up observations that might characterize its discoveries.

The other approach — making hyperaccurate measurements of a star's position — is the goal of NASA's US$1.85-billion Space Interferometry Mission, or SIM, a bone of contention between the agency, which has sought to cut its funding, and its fans in the astronomical community. Zlatan Tsvetanov, the NASA programme scientist for SIM, says the astro-comb technique does not replace the SIM technique "by any means". But a forthcoming exoplanet report, commissioned by a National Science Foundation advisory committee, says that the number-one priority is finding another Earth in the most cost-effective way. If discovering Earth-sized planets with a technology that cost less than a million dollars really works, it may be hard for billion-dollar space missions to compete. ∎

Eric Hand

SNAPSHOT
Flirty in pink

Lustily trashing some hollow-stemmed matting grass, this pink Amazonian river dolphin (*Inia geoffrensis*) is working it down the waterway in Mamirauá, a flooded rainforest reserve in Brazil. It's almost an exculsively male practice, this brandishing of grasses, sticks or clay. It is thought to be a courtship ritual, although no researcher has seen the dolphins mating, despite more than 13,000 hours of observation. If so, it places dolphins, with humans, chimps and birds, among the few animals that use objects as a flirting device.
Anna Petherick

Back to basics for HIV vaccine development

The US programme to develop an HIV vaccine will shift funding back towards basic research and away from clinical trials of candidate vaccines, officials at the National Institutes of Health announced last week.

The decision follows the discontinuation last year of trials for Merck's candidate HIV vaccine, in which participants showed higher rates of HIV infection than the unvaccinated population. Merck's was one of only three vaccines to make it to large-scale clinical trials. A second had already failed, and there are doubts about the ability of the third to stimulate the immune system.

The vaccine pipeline needs refilling, National Institute of Allergy and Infectious Diseases director Anthony Fauci said on 25 March. "We need to turn the knob in the direction of discovery — that is unambiguous."
See Editorial, page 503.

US medical school to require master's degree

The Scripps Research Institute and its clinical partners are forming the first US medical school that will require applicants to have a master's degree. Scripps, which is based in La Jolla, California, expects to enrol its first 40 students in 2010.

"The master's will be in translational medicine," says Scripps cardiologist Eric Topol, the school's founding dean. "Our mission is to encourage a new genre of physician–investigator."

The medical training will take place in existing facilities, but Scripps is seeking philanthropic support for much of the school's $150-million cost, Topol says.

Many US medical schools offer combined MD–PhD programmes for some students, but Scripps is the first to make a master's degree a requirement.

Too much haste in US drug approval?

The US Food and Drug Administration (FDA) may be issuing shaky last-minute drug approvals, according to a new study.

In an article in the *New England Journal of*

Tasting a geyser

The geyser of ice and water vapour erupting from the south pole of Enceladus, a tiny moon of Saturn, contains complex organic molecules, NASA scientists have discovered.

The Cassini spacecraft flew through the plume on 12 March and got its first good taste of the jet's chemistry — which is unexpectedly similar to that of a comet.

The flyby results, reported on 26 March, also include thermal images of the four long cracks from which the plume emanates. At their hottest, the cracks were 93 °C warmer than the rest of the moon, indicating underground reservoirs of liquid water. This all suggests that Enceladus formed differently from, and probably later than, the rest of the Saturn system.
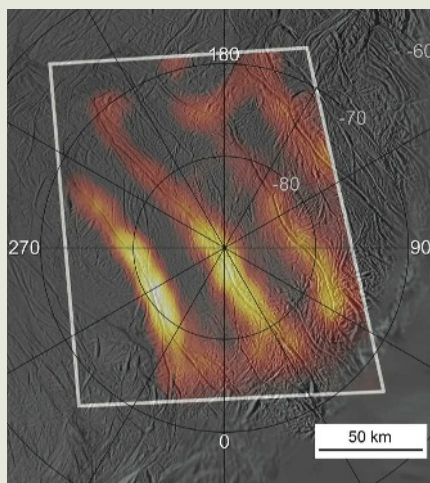
NASA/JPL/SSI

Medicine ((D. Carpenter *et al. N. Engl. J. Med.* 358, 1354–1361; 2008), Daniel Carpenter and his colleagues at Harvard University studied 313 drugs approved for market between 1993 — when drug companies began paying for reviews, and penalties came in for the agency missing deadlines — and 2004.

They found that 14% of drugs approved in the two months before deadlines had serious safety problems requiring 'black box' label warnings, market withdrawals, or both, compared with 3.2% approved earlier in the process. The FDA is disputing Carpenter's figures.

Italy to help Kenya develop space programme

The Italian Space Agency, which operates a tracking station in Kenya, will give the country access to geospatial data and training in satellite systems.

Under an agreement signed on 28 March, six Kenyan scientists and engineers will travel to Rome to help build MIOSAT, an Earth-observation satellite. Italy will also set up a centre in Kenya to give scientists there access to Earth-observing data. The project is expected to cost the Italian government some €200,000 (US$315,000).

The agreement is a major step forward

Italy runs a tracking station in Malindi, Kenya.

for Kenya's nascent space programme, says Zachary Mwaura, the country's secretary for defence. "This is something that we have really, really been looking forward to," he says.

Muslim nations raise status of science

Science is being taken more seriously in the Organisation of the Islamic Conference (OIC), a group of 57 states.

The OIC's standing committee on science and technology (COMSTECH) is set to announce this week that only countries

committed to donating funds will be allowed to join the committee. Payment will be in proportion to gross domestic product, at a rate set by the OIC.

COMSTECH will use the money to launch programmes for its members in areas such as biotechnology, development of pharmaceuticals from indigenous raw materials, and science policy.

Until now, COMSTECH has been open to all OIC members, but few gave any money. Last year, only US$2.59 million was taken from 13 states — most of it from Pakistan — which was used to support library services.

Atta-ur-Rahman, who heads both COMSTECH and Pakistan's Higher Education Commission, expects up to 15 countries to join this year, with more joining in the future. "Most of the Muslim countries don't have national science policies based on their local needs and available resources," he says. "COMSTECH will help them."

Correction
Our News story on the Quake-Catcher Network (*Nature* 452, 397; 2008) should have noted that the concept originated with Elizabeth Cochran, a geophysicist at the University of California, Riverside, and that the project is now an ongoing collaboration between her and seismologist Jesse Lawrence of Stanford University in California.

Hazy reasoning behind clean air

Science alone can't determine how regulations are written, argues **David Goldston**.

Last month, *The Washington Post* reported that President George W. Bush had personally intervened to weaken new regulations to control smog just as they were about to be announced by the Environmental Protection Agency (EPA). In response, advocates of tighter standards predictably charged that the president had overturned a scientific judgement. Carol Browner, who headed the EPA under President Bill Clinton, put the matter starkly, telling the *Post* that the Clean Air Act creates "a moral and ethical commitment that we're going to let the science tell us what to do".

But does it? This conceit that science alone should and can dictate clean-air standards is propagated by political figures of all stripes and often by scientists themselves. Politicians always want to argue that any regulatory measure they are supporting is the only one justified by science because doing so makes their position sound objective and above the political fray. That's especially true in today's polarized environment, when claiming to have science on your side may be the only line of argument that can reach someone who doesn't share your ideological persuasion.

In reality, though, regulatory decisions involve policy judgements as well as scientific determinations, and the science is often uncertain. The Clean Air Act explicitly leaves decisions to the "judgment of the administrator" of the EPA (a presidential appointee), who is advised by, among others, a scientific panel. Contending that standards are based solely on science conflates policy and science questions, muddying the debate and putting scientists needlessly in the line of fire.

So what's really at issue in the case of the smog rules? The rule the president changed sets what is known as the secondary standard for allowable concentrations of ground-level ozone, the main component of smog. Under the law, the secondary standard is designed to "protect the public welfare" by preventing damage to crops, natural vegetation and anything else other than "public health", which is covered by the primary standard.

The EPA's 24-member scientific advisory panel weighed in on two critical questions concerning the secondary standard: over what time period should ozone be measured, and what should the permissible level of ozone be? The first question may sound like a technicality, but it affects which

PARTY OF ONE

areas turn out to violate the standard because ozone levels can vary significantly within a given day. For example, if being above the allowable level for even one hour on one day is a violation, then places may be required to clean up their air that wouldn't if the rule forbade exceeding that same standard over an eight-hour average on a single day.

The time period is supposed to be based on biology. If being exposed to ozone for even one hour can have a substantial impact, then that should be the time period used. The advisory committee said unanimously and unequivocally that plants are affected by ozone through cumulative exposure over an entire growing season — so the existing eight-hour period, which is based on how ozone affects humans, is inappropriate. The president left the eight-hour period in place, which clearly runs counter to the science. For that piece of the rule, he should be considered guilty as charged.

But the other piece is deciding what level of ozone should be permitted, and that cannot be determined solely by science for two reasons. First, deciding what level of damage constitutes a threat to "public welfare" inherently is not a scientific question. Scientists may be able to describe the damage that could result from a given level of ozone, but the decision that such damage is so great that it must be prevented is a policy matter.

Second, the EPA's science panel found that "quantitative evidence linking specific ozone concentrations to specific vegetation/ecological effects must … be characterized as having high uncertainties." What to do in the face of uncertainty is a policy question, not a scientific question. So although the advisory panel

unanimously recommended a specific range of ozone standards, a number within that range can hardly be seen as the only justifiable standard under the law. Indeed, the EPA's own science staff had recommended a slightly different range. Critics are free to attack the number chosen by the president, which will keep some rural counties in compliance with clean-air rules. What they cannot legitimately argue is that the president's selection runs counter to the science. The debate is about what kinds of damage harm the public welfare and what kinds of uncertainty can be tolerated as a basis for decision-making.

The debate over the new ozone standards is just beginning, but the detrimental impact of confusing science with policy can be seen by looking back at what happened in 1997, when the EPA last changed the ozone rules. The fight then was over the primary ozone standard, the one designed to protect public health. The EPA proposed tightening the standard, and Browner (then EPA's chief) repeatedly argued that the decision was dictated by the science.

As a congressional staffer, I fought for the EPA proposal and I still support it. But what the science actually demonstrated was that for a given level of ozone, there are a predictable number of excess hospital admissions from aggravated respiratory conditions. At the time, there was little indication that ozone caused chronic health problems or deaths. Therefore the policy issue was: "How many hospital admissions are acceptable?" Needless to say, no politician was interested in engaging in that debate. The members of the EPA's science advisory panel at the time were split over what standard to suggest, but agreed that the number was a "policy call", not a scientific question. The science in no way told Browner exactly what to do.

All this quickly got lost in what became a prolonged and highly acrimonious debate between supporters and opponents of the new rule, in which each side accused the other of using poor science. This was bad for policy because the question of how to decide on an acceptable level of protection never got raised, never mind discussed. And it was bad for science because accusations of poor science conducted in the service of political goals can only raise distrust and confusion about the scientific enterprise.

The 1997 ozone fight, even more clearly than the 2008 rerun, was a case of a policy debate masquerading as a science debate. In such instances, scientists ought to be busy ripping off the policymakers' masks, not donning them. ∎

David Goldston is a visiting lecturer at Harvard University's Center for the Environment. Reach him at partyofonecolumn@gmail.com.

An ultra-low-energy passive house project in Lindås, Sweden

M. WALL

Architects of a low-energy future

Low- and zero-energy buildings could have a huge impact on energy use and carbon emissions. We have the technologies, but if they are to mitigate climate change, green-building design must hit the mass market, says **Declan Butler**.

"It felt surreal," says Karsten Voss, thinking back to January 2008 and the winter meeting of the 50,000-member international heating, cooling and ventilation research association, ASHRAE. "Here we were sitting talking about zero-energy buildings, one of the biggest topics on the programme, inside a hotel that had no proper glazing or insulation — while it was –10 °C outside."

It was a classic illustration of both the challenge and the promise of green architecture, says Voss, who should know. An architectural engineer at the University of Wuppertal in Germany, he also heads assessment of ultra-low-energy building demonstration projects built under the country's national programme on research for energy-optimized construction (EnOB). Germany has pioneered such research, thanks to its long-standing interest in the environment, and has some of the toughest energy-efficiency regulations around.

Buildings worldwide account for as much as 45% of energy consumption, and a similar share of greenhouse-gas emissions. That makes buildings the biggest single contributor to anthropogenic climate change — a worse offender than all the world's cars and trucks

put together. But it also makes the design of energy-efficient buildings a "monumental but essential task" in the effort to mitigate climate change, according to a research roadmap published last November by the research committee of the US Green Building Council. Indeed, the latest report from the Intergovernmental Panel on Climate Change (IPCC) estimates that improvements in the energy efficiency of buildings could potentially reduce projected global carbon emissions up to 29% by 2020, and up to 40% by 2030. Moreover, the IPCC's estimates are deliberately conservative, based on a pessimistic view of how rapidly the building industry can reform its practices. If the reforms could somehow be accelerated, the reductions in fossil-fuel energy needs could be that much more dramatic.

Such a speed-up is certainly possible — in principle. Voss and his fellow 'green-building researchers' have already developed and tested many of the necessary tools and strategies, to the point where they can now build homes, offices and other buildings that use 80–90% less

> **"We don't think about energy in the design phase."**
> — George Jeronimidis

energy than existing buildings. The most efficient of these structures are almost completely 'passive', meaning they require very little, if any, traditional heating or air-conditioning. Yet the overall comfort they provide is, if anything, superior to existing buildings. Nor is there necessarily a cost penalty: these ultra-energy-efficient buildings are often no more expensive to build than conventional structures, and work out far cheaper if energy bills during their occupation are taken into account.

In practice, however, getting the construction industry to change its ways is a daunting prospect. Especially in the developed world, more than a century of cheap energy has divorced the architecture of buildings from energy considerations, and from their environment. Architects typically design whatever their client wants, mostly on the basis of cost or aesthetics, and then pass the design to engineers who have to make the building work and be comfortable — typically by bolting on energy-intensive air-conditioning, heating, artificial lighting and whatever else it takes. As a result, most builders and architects

today haven't got a clue how to design buildings with ultra-low carbon emissions.

And even if they did, there is the sheer scale and diversity of the trillion-dollar worldwide building industry — a vast labyrinth that involves as many different professions and sectors as there are sorts of wallpaper, plus an almost unlimited number of types of buildings, regulations, norms, materials and environments. In the face of all that, "the major impediments to increase energy efficiency in the building sector are institutional barriers and market failures rather than technical problems," notes a report on green buildings published last year by the United Nations Environmental Programme. "Even if high-tech competence exists in most countries, the institutional and economic conditions have hindered the technical competence to be effectively applied in day-to-day design, construction and operation of buildings."

Change is in the air, however. Green buildings, once the preserve of ardent environmentalists, are going mainstream. Architect Edward Mazria of Architecture 2030, a non-profit organization in Santa Fe, New Mexico, set a challenge in 2006 — a voluntary commitment to reduce the carbon consumption of all new and retrofitted buildings by 50% immediately, and to meet further reduction milestones to become carbon

neutral by 2030. These targets can already largely be met using existing technology, argues Mazria, who has been practising sustainable architecture for 30 years. And indeed, the challenge has been endorsed by the US Conference of Mayors, several US cities, the American Institute of Architects and more than 400 other professional organizations.

Meanwhile, the US Energy Independence and Security Act, which came into force in December last year, requires all federal buildings to meet stepwise goals almost identical to those of Architecture 2030's challenge. The act also introduces tougher energy targets for both commercial and residential buildings. In Europe and elsewhere, similar, or even tougher, targets are often being made legally binding on many buildings.

There is a "sea change" underway, says Gregor Henze, an architectural engineer at the University of Nebraska-Lincoln in Omaha. Carbon emissions weren't even on the radar of most architects five years ago, he says, adding that the number of architect firms, construction companies and other members of the US Green Building Council is "exploding" — it now has 13,500 member organizations, and 91,000 members, up some ten-fold since 2000.

One driver is growing public concern over spiralling fuel prices and climate change, says Ken Baker, chief economist of the American Institute of Architects. The institute surveyed house purchasers' expectations at the end of 2007 and found that, despite the subprime mortgage crisis, most were willing to pay more for a house that would use less energy and be kinder on the planet. "Consumers are becoming increasingly aware of energy-efficient options," says Baker, "and they are requesting that architects incorporate them into the design and remodelling of their homes."

> **"Consumers are increasingly aware of energy-efficient options."**
> — Ken Baker

Innovate and renovate

Meanwhile, green-building researchers are working hard to expand those options. The most obvious and urgent priority is to take energy into account during retrofitting. Although new construction in Europe and the United States annually amounts to around 1% of the existing building stock, about twice that many structures are renovated every year. Given the typical life cycle of buildings, each renovation that uses inadequate insulation or poorly glazed windows is an opportunity for lowering energy demand that is lost for decades. One demonstration project carried out at the Jülich Research Centre in Germany shows what can be done with energy-efficient design strategies. Although the gains are not as impressive as those with new buildings, the renovations nonetheless cut the test building's total primary energy needs in half, from 1,235 kilowatt hours per square metre per year ($kWh/m^2/year$) to 600 $kWh/m^2/year$.

Ultimately, however, the biggest pay-offs will come from new buildings, where ultra-low-energy use can be designed in from the beginning. This is particularly important for the developing world, and urban areas, where so much new construction is expected in coming decades. The goal in such ultra-low-energy design is to make buildings as 'passive' as possible, meaning they can satisfy most of their occupants' heating, cooling, and lighting needs from the outside environment — and then maintain the occupants' comfort as the outside cycles from day to night, summer to winter.

More research is still needed, for example, to develop cheap glazing and insulation materials that can adapt themselves to changes in lighting and heating loads over the course of a day, or a season. Research is also vital to adapt existing solutions to different sorts of buildings or environments.

Nonetheless, many of the technologies are already mature. Almost two decades of

WWW.PASSIV.DE

A 1950s building in Frankfurt, Germany before (left) and after (right) refurbishment. Thermograms (top) show the reduction in heat loss — energy demand dropped 90%.

The Centre for Environmentally Conscious Construction in Kassel, Germany, leads by example.

research, mainly in Europe, has resulted in designs for fully passive houses that require primary energy needs of 15–50 kWh/m^2/year — compared with current-generation European houses that typically require 160–300 kWh/m^2/year — and need only supplementary heating or cooling, in tiny amounts, or just on the coldest and hottest days.

Such efficiency levels make the buildings amenable to being made zero-carbon by meeting their low energy needs, either from small, local renewable sources, or using grid electricity generated from renewable sources.

"You can do a lot with existing technology", says George Jeronimidis, an engineer who heads the Centre for Biomimetics at the University of Reading, UK, and who is developing next-generation building materials that adapt their shape, such as vents that open and close in response to humidity levels. "We are missing a lot of opportunities because we don't think about energy in the design phase of a building."

Holistic house-building

This does require a different mindset, says Maria Wall, a green-building researcher at the University of Lund in Sweden who was involved in building passive terrace houses in Lindås (see photo, page 520). "Getting architects and engineers sitting down together from the outset is critical," she says. "One has to consider the building as a whole, and in the context of its environment." But new software is helping architects and engineers do just that, says Andreas Wagner, a building scientist who heads the architecture faculty at the University of Karlsruhe in Germany, and who is also a member of the design and monitoring team within the EnOB programme.

The software provides a "common language" that allows architects and engineers to jointly quantify the energy implications of design decisions. Using databases and models of average temperatures, sunlight and solar radiation from different north–south orientations, it allows them to pick from a supermarket of passive design strategies, techniques and materials to adapt designs to the local conditions. "We

are now moving on from research and prototypes into the building market," says Wagner.

That market is most mature for residential homes, which account for three-quarters of the energy consumption of the building sector. Passive houses currently make up only 2% of new buildings in Germany. But market research suggests that fraction will climb quickly as the necessary competences spread, and as consumers begin to take advantage of low-interest loans and other recent government incentives. Passive houses are also spreading fast in Austria, Switzerland and other European countries.

Passive-house design uses extreme insulation of the building envelope — including triple-glazed windows, often filled with an inert gas — in particular, in an effort to eliminate what engineers call 'thermal bridges'. These are fault lines, typically at the window frames and at the intersection of floors and walls, where large, unwanted energy transfers occur with the outside. The result is an airtight 'skin' that

prevents energy leakage. Thermograms of buildings then show impressive gains in 'negawatts' (see photos, page 521).

With insulation like that, the building can get its heating from the solar gains through glazing as well as through waste heat from appliances and even our bodies.

Another key technique for temperature control in passive houses is at first counterintuitive: simply let fresh air in from the outside. A pump draws fresh air through a grid of pipes buried several metres underground, where the temperature is relatively constant throughout the year, 10–14 °C in the United Kingdom, for example. When this fresh air arrives at the house, its temperature has already been modulated — warmed up or cooled down by the ground, depending on the season. Then, in a second trick for heating, the incoming fresh air is put through a heat exchanger to recover 80% of the heat from the warmer stale air being expelled from the building. This system of air-based cooling and ventilation not only saves energy by recycling heat, but vastly improves air quality.

Many of these same tricks can be used for office buildings and other large structures. However, these buildings tend to be more complex than residential houses, with many more interactions between different interior zones and higher heat loads from staff and machinery. So it is considerably more difficult to make them completely passive. Then, too, research in this area is relatively recent, even in Germany. Still, EnOB recently published results from

Energy-efficient homes in Lenoir City, Tennessee (left), and in Oberdorf, Switzerland.

the construction of 25 prototype passive office and other large buildings, and found that most kept energy for heating, lighting and cooling below 100 kWh/m²/year. "That's a factor of four smaller than existing equivalent buildings in Germany, and of six than in the United States," says Wagner. Three of the buildings reached energy consumptions of less than 50 kWh/m²/year.

The EnOB study tested multiple concepts, including passive cooling. One such concept is night ventilation, which is increasingly being deployed in large buildings worldwide. The idea is to open automatically the skylights during the night, when the air in the building is warmer than the outside air. A natural breeze will then release the hot air out the top while drawing in cooler outside air from ventilation grills in the facade.

Cool control

Another concept tested in the study was to circulate water in boreholes drilled up to 100 metres down, and then pass the water in closed circuits through concrete slabs in the ceilings of the building. The water was a relatively warm 17 °C. But there was so much of it that the cooling effect was comparable to that achieved by air-conditioning compressors. "We use the environment, the ground, as a heat sink," says Wagner.

Many new German houses now use such slab cooling, or 'thermally activated building systems', which were pioneered in Switzerland. The buildings require only very small amounts of energy; all they need is a small pump. A multi-million-euro renovation of the United Nations offices in Geneva, Switzerland, will use a similar system, drawing water from Lake Geneva.

These systems will work in even the hottest areas of California, Wagner maintains, although a small amount of conventional air conditioning might have to kick in during the late afternoon to keep temperatures below what people have become accustomed to. "But it would reduce most of the current needs for air conditioning."

One problem is that the range of daytime temperatures achieved by passive cooling and ventilation are still less predictable than what conventional systems can deliver, in particular in large buildings; careful modelling and scaling are required to cover the possible extremes. As a result, risk-averse architects and engineers worried about potentially unhappy clients tend to stick with the safest option, and sell buildings that use central heating and air conditioning.

"The need is to build more demonstration passive buildings, to show everyone from carpenters to house purchasers that it is not such a strange thing to build an energy-efficient house," says Wall, "and that not only do these buildings work, but they are actually superior to conventional buildings."

Fully passive designs can be adapted to hotter climates, although much less research has been done for those areas. In some cases, it works out cheaper to use a strategy based on renewable energy. That's the tack being taken by Jeff Christian, head of the Buildings Technology Center at the Oak Ridge National Laboratory in Tennessee, to design cheap, energy-efficient homes for low-income families.

In Tennessee, he has to deal with not only heat, but also with high humidity. Underground passive ventilation systems bring in too much moisture for such environments. And although there are possible solutions, such as solar-driven dessication, this is currently too expensive.

Nonetheless, says Christian, low-energy buildings for the mass market are already doable there. His focus is what he describes as "a little bit of solar, and a lot of energy efficiency" — that is, using better insulation and using solar panels. His building in Lenoir City, Tennessee (see photo, page 522), also maximizes cooling through natural cross-ventilation. The solar panels cost $20,000 upfront, he says, but he cut costs by mass-producing wall panels and other parts of the building.

For all the progress, though, Christian, for one, is convinced that cheap, low-energy houses will take off in the United States only if the government wields both carrot and stick. "The financial incentives we need to drive this are not in place," he says, adding that tougher and compulsory energy requirements on buildings "is the only way we will get industry to do the research and development to get there".

Yes, he says, green buildings can potentially be built as cheaply as their conventional counterparts, if not more so. But those psychological and institutional barriers are still very real. If engineers, architects and builders have to struggle with, for example, unfamiliar green-architecture techniques, they will incur a heavy 'transaction cost' in the form of confusion, mistakes and delay. And if they have to add renewable energy, they will saddle the project with higher capital construction costs — which clients in the large rental sector will have little interest in paying, as it is the tenants who will reap the rewards in the form of lower energy bills.

The mass-production of wall panels and other parts of the building advocated by Christian is also seen as a promising avenue in Europe. Cost can be reduced further by avoiding one-off designs and constructing an identical series of buildings. "The passive terrace houses we have here in Wuppertal are identical to those in [the German cities of] Wiesbaden or Hannover," says Voss. "The only difference is the colour."

> ### "The financial incentives we need to drive this are not in place."
> — Jeff Christian

Nonetheless, it is still an open question how quickly passive architecture strategies, combined with renewable energy sources, will become the mainstream of construction. The good news is that the interest is there. Witness the turnout of 30,000 delegates in Chicago last November at the "Greenbuild" conference — which was opened by former US president Bill Clinton. On the last day of the conference, Maria Atkinson, head of sustainability at Lend Lease, a multinational real-estate company, threw out a challenge that should please Voss. "The hotels of Boston have 12 months to step up to the challenge of ensuring they are green for Greenbuild 2008, because this year's delegates will be demanding green hotels as part of their commitment to green buildings." ■

Declan Butler is a senior reporter at *Nature*, based in France.

The world's largest 'passive' office building in Ulm, Germany.

ENERGON

TRULY GROSS ANATOMY

Sometimes it is necessary to immerse yourself in a subject. **Erik Vance** meets a woman whose research takes her deep — waist-deep — into cetacean anatomy.

If you work as a toll collector at the Lincoln Tunnel in New York City, you might think you've seen it all. Until the day a woman drives past your booth with a dead dolphin stuffed in her hatchback.

"I had to fold the front seat over on the passenger's side and shove the bottlenose dolphin's face out of the front passenger window to fit the thing in the car," says Joy Reidenberg, almost losing her breath in laughter at the memory. "And so coming in through the Lincoln Tunnel, they're saying 'What do you have as a passenger?'"

Almost all conversations with Reidenberg, a specialist in cetacean anatomy at the Mount Sinai School of Medicine in New York, end with her deep, hearty laughter. In an age where molecular-level work is king and medical students rarely touch dead bodies, she is one of a handful of researchers left who still embrace the title 'anatomist'. Although colleagues describe her as one of a dying breed, she regularly publishes research that breaks new ground — and maintains that anatomy is as important today as it was in the days of Aristotle.

"But this is not for the weak stomach," she says in her distinctive New York accent. "Breathe through your mouth and hope the wind is strong."

To obtain specimens for her research, Reidenberg has to bring back parts of cetaceans stranded on beaches from Florida to Maine. This means all-day marathons that are a cross between a Jacques Cousteau film and a night at the graveyard. "To bring back a larynx the size of this table," she says, thumping on a conference room table, "takes six people, a tug-of-war, and maybe a backhoe and a crane."

Reidenberg has extracted body parts such as fins, brains and tracheas from all manner of cetaceans, including once from a massive blue whale. This may sound simple, but body parts can weigh tonnes and carcasses are often hard to access. At times she has been chased by sharks, lowered by crane into intestines as deep as a swimming pool, and even pinned against a boat by the errant rolling head of a pilot whale.

Yet she finds plenty of eager volunteers to help. "She was, and is, incredibly passionate about this type of anatomy. And that's pretty infectious," says Armand Balboni, a former student who is now a postdoctoral researcher at Mount Sinai. Her gross anatomy classes are hugely popular — although it's not clear whether this is because of her engaging personality or her habit of bringing whale flippers and sea turtles to class.

But Reidenberg is more than just a popular teacher with an overflowing freezer. She holds academic appointments at the Smithsonian Institution, the Woods Hole Oceanographic Institution and the New York Consortium of Evolutionary Primatology. She has dissected more than 140 cetaceans of various sizes, representing 19 different genera. And she publishes regularly across a wide spectrum of topics.

Much of her work examines the cetacean throat — including how breathing, eating and sound genera-

Joy Reidenberg finds that anatomical studies on whales can involve efforts as huge as the mammals themselves.

> "This is not for the weak stomach."
> — Joy Reidenberg

tion work together in an underwater environment. For instance, the mechanisms by which some baleen whales release clouds of tiny air bubbles to confuse prey have long been a mystery. By comparing seven humpback whale carcasses Reidenberg discovered some odd morphology in the epiglottis, a structure that helps separate eating and breathing. She deduced that it is specially formed to flip back and forth, channelling air either to the blowhole or to the mouth, where it would be forced through the baleen as a bubble cloud. This not only answers a question of mechanics, but it means that bubble clouds — which whales may use daily in colder waters — are valuable enough to risk water from the mouth entering the trachea and asphyxiating the animal (J. S. Reidenberg & J. T. Laitman *Anat. Rec.* **290,** 569–580; 2007).

Reidenberg has also looked into the mystery of sound generation in cetaceans. Vocal folds don't seem to exist in whales, and many researchers have assumed that most

noises come from the nasal area. But that is because vocal folds in land animals run perpendicular to air flow. Reidenberg discovered a structure in cetacean larynges that runs parallel to air flow, thereby not blocking the airway when it vibrates (J. S. Reidenberg & J. T. Laitman *Anat. Rec.* **290,** 745–759; 2007). She discovered a similar structure in hippos, the closest living relative to cetaceans.

Evolution of the cetacean voice is still mysterious because the soft tissues of the throat don't fossilize well. But to Reidenberg, her discovery reinforces the idea that whenever possible, nature finds solutions using structures at hand, rather than inventing something new. In the case of whales, low-frequency sound travels well underwater. To make such sounds, you need longer folds. Long folds can't fit perpendicularly across the airway, so nature turned them 90° for more elbow room.

The fear factor

Reidenberg's closest associate and mentor, Jeff Laitman of Mount Sinai, co-authored both studies. But as a department head he doesn't always share her enthusiasm. "The two words that have usually sent me into near heart attack have been the words 'mass stranding'. To her, this is one of the greatest excitements in the world," he says. "I, on the other hand, have the wonders of explaining to the institutions, the boards, the loading dock, the security people the wonderful material that comes in — trying to stand there with a straight face and say, 'Odour? What odour?'"

With a lab overlooking Central Park, taking up some of the world's most expensive real estate, Laitman says that he cringes every time Reidenberg excitedly calls to say that she has contracted a flatbed truck to bring back some giant specimen. "She's out there and my thoughts are, 'Oh Lord, please let it be something in which the larynx is less than five feet'," he says.

There are upsides to the job. For instance, Reidenberg rarely has to open her carry-on luggage once inspectors learn what's inside. And then there was the time she was asked by an airline whether her travel was because of a sudden death. Not thinking, she said it was. She flew to North Carolina under a special rate, presuming she was in mourning for a dead sperm whale.

As a child in suburban Connecticut, Reidenberg says that she was always more interested in trucks and road kill than dolls. Once, at about age eight, she decided one of her few dolls needed a fur coat. Rather than asking her mother for one, she went out and found a dead chipmunk. She flayed it and was drying the hide when, to her horror, a raccoon took her prize.

Today her road kill requires a little more work. In addition to trucks and cranes, a large whale necropsy requires perhaps a dozen people with specialized hooks and knives, which have often been often confiscated from whaling ships that wander into US waters. Dead whales are categorized into one of five codes, based on the level of decomposition. Code 1 is freshly dead, or as Reidenberg says, "like walking into a butcher shop". Code 5 is essentially a dried skeleton. The most common is Code 3, a terrible middle ground where the carcass still contains useful scientific material, but can be best described as putrid. At this stage, the dark red meat has turned tan and is peeling from the body, and

COURTESY OF J. REIDENBERG

Not in my back yard: this fin whale is in immaculate condition compared with some decomposing cetaceans that wash up.

organs have liquefied into an indecipherable mush.

Many of the best dissections happen in the winter when the animals are preserved by the cold. But the thin latex gloves Reidenberg wears mean that she must intermittently bury her hands in the carcass to stay warm enough to feel her tools. Yet she will happily work for 12 hours extracting a brain that isn't even for her own research.

Lori Marino, a neuroanatomist at Emory University in Atlanta, Georgia, who works with Reidenberg, calls her a "scalpel and table" anatomist. In truth, she is a hammer, chisel, axe, flensing knife and hand-saw anatomist. (She spurns chainsaws for their "fling" factor.) Marino says this gives her a crucial perspective over the entire body that few scientists have in this era of specialization.

Together, Reidenberg and Laitman argue that researchers cannot understand molecular biology or gene expression without better understanding the form and function of anatomy. Further, they say that putting the human body in context requires understanding how it relates to other mammals. Yet anatomy remains a hard field to lure students to, and even harder to find funding in.

Although Reidenberg's work has informed researchers who study human throat maladies such as acid reflux, she admits that it is not the kind of thing that interests important backers such as the National Institutes of Health. In fact it's so hard to get gross-anatomy funding that many medical schools don't expect their professors to publish at all. According to the American Association of Anatomists and various anatomy department heads, this leads to a vicious circle with fewer academics seeing a future in anatomy and then even less funding. "Where is the next generation of anatomists coming from?" asks Reidenberg. "This is a big problem."

She is trying to do her part in her classroom. But if you have a question about a lecture, you may not find her during office hours. That is, unless you are handy with an axe and don't mind the smell. ∎

Erik Vance is a freelance science writer in Berkeley, California.

> **"The two words that send me into near heart attack are 'mass stranding'."**
> — Jeff Laitman
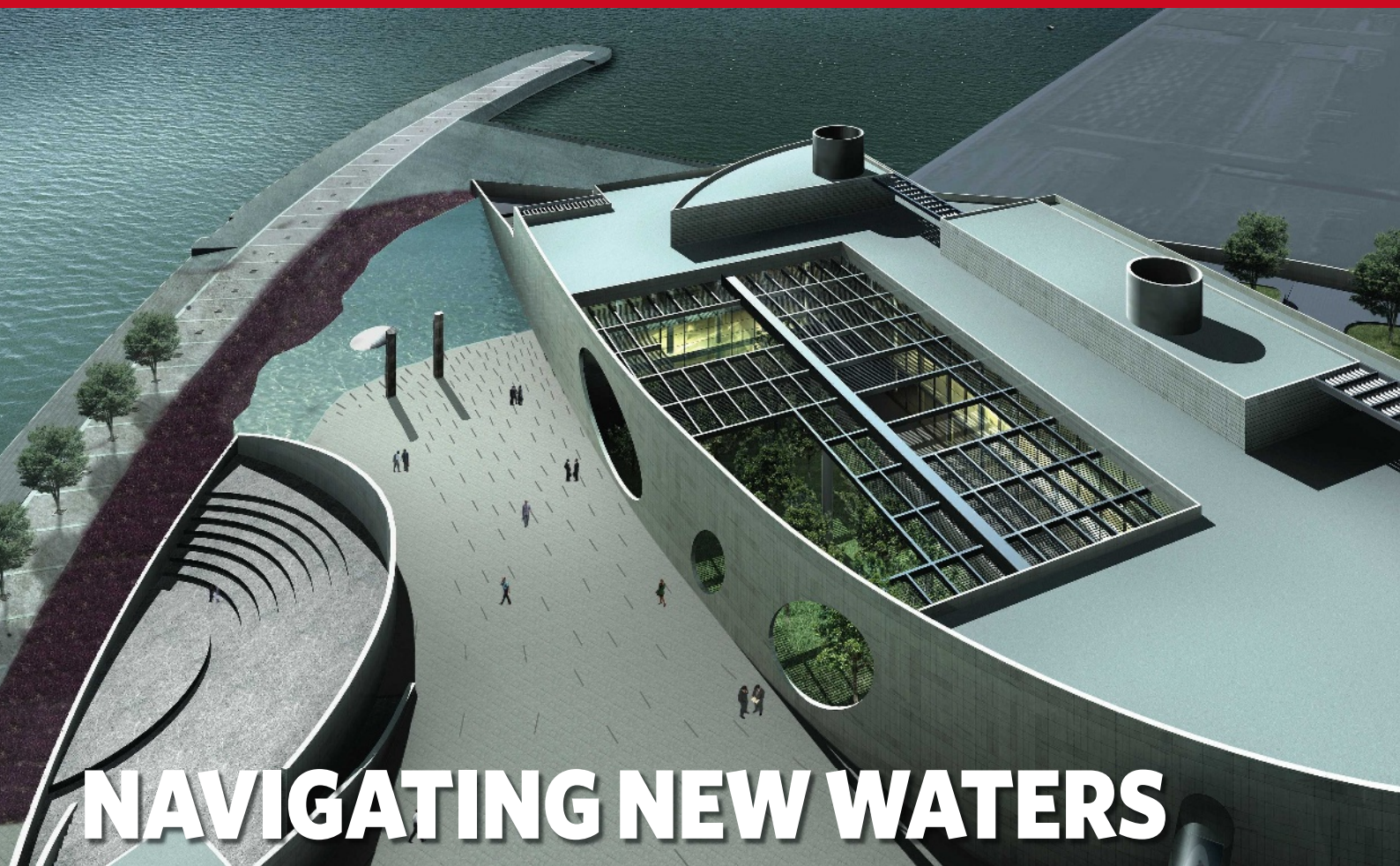
NAVIGATING NEW WATERS

Portugal's spending on research is near the lowest in western Europe. Can a single-minded lady with half a billion euros change things, asks **Alison Abbott**.

Three years ago, everyone seemed to want neuroscientist Zachery Mainen. Janelia Farm Research Center, the Howard Hughes Medical Institute's prestigious new hub in Ashburn, Virginia, had just made a generous offer to recruit him and his frontier studies on decision-making. His current employer, Cold Spring Harbor Laboratory in New York, made a more tempting counter offer. But it was a third, unexpected proposal that he couldn't refuse. Even though the research centre in question did not yet exist. And even though it was in Portugal.

Never in his most fanciful moments had Mainen imagined shifting his career to this small, relatively poor corner of Europe. Yet last July he became one of the first scientists to be signed up by the Champalimaud Foundation Research Centre in Lisbon. When completed in 2010, the centre will be the most grandiose and expensive life-sciences research project the country has even seen.

The scheme has been master-minded almost single-handedly by a forceful lady named Leonor Beleza. Beleza knew little about science until 2004, when she found herself named president of a half-billion-euro endowment for biomedicine left by Portugal's richest man, António de Sommer Champalimaud, on

his death. Beleza spent a year or so consulting with the world's scientific élite before deciding how to spend it — lavishly, as it turned out. The Champalimaud Foundation Research Centre will have everything money can buy: a striking US$120-million building designed by renowned architects, top-of-the-range equipment and at least 300 scientists. It will also have a waterfront location on the very spot from which Vasco da Gama and other fifteenth-century navigators departed to discover the unknown world.

What convinced Mainen to move there was the opportunity to pursue his research — on how animals use experience and sensory input to make decisions — in any direction he chose, and without undue pressure to secure additional funding. "It's got the backing to be a world-class research institute," he says from his temporary laboratory at the Gulbenkian Institute of Science in Oeiras, where he is working until the new building is finished.

It is as bold a project as can be imagined in a country that for some time has had one of the lowest investments in science and research in the European Union outside the former communist member states. In the past few

years, the Portuguese science base has grown and modernized rapidly, and spending has risen, says research minister José Mariano Gago. Even so, in 2005, Portugal spent 0.8% of its gross domestic product on research and development, compared with an average in the European Union of 1.8%. Some hope that the unprecedented injection of money from Champalimaud will quicken efforts to transform the country's research. "This new institute will make a huge difference to life sciences in Portugal," says António Coutinho, director of the Gulbenkian Institute.

> "This new institute will make a huge difference to life sciences in Portugal."
> — António Coutinho

Generous tycoon

Champalimaud began building up his fabulous fortune in the 1930s with a cement factory. Portugal, a dictatorship since 1926, was a friendly country for the determined capitalist at the time. But after the 1974 'Carnation revolution', the new socialist regime forced Champalimaud nearly penniless into exile. Undaunted, he rebuilt his wealth in Brazil and, with the maturing of Portugal's democracy, Champalimaud was able to return home in 1992 and buy back many of his former companies.

Champalimaud was a very private man and made his own decisions. It was only after his death that his relatives found out he had earmarked one-quarter of his wealth for biomedical research. But on the day Champalimaud signed his will, he called Beleza out of the blue to ask in confidence whether she would agree to be president of the planned foundation. "I was flattered — but really surprised," she says. "I had only met Champalimaud fleetingly on a few occasions over the years and was not aware he had particularly noticed me." Attracted by the philanthropic opportunity, Beleza said yes. She heard nothing more until Champalimaud died four years later.

Beleza, a carefully elegant and engaging lady nearing 60, trained as a lawyer but spent most of her professional life in politics. Known for her dogged determination, she was once considered Portugal's Margaret Thatcher. But she dropped from the limelight after she was implicated in a scandal involving the distribution of blood contaminated with HIV during her stint as health minister in the late 1980s. More than 120 people with haemophilia had become infected by blood products imported from Austria.

Uncharted territory

Beleza pulled out of politics altogether when she took up the reins of the Champalimaud Foundation and had to decide what to do with the sudden fortune. Champalimaud had left no guidance on how the money should be spent other than that it should be for biomedical research. Seeking direction, Beleza took herself on a brisk tour of the scientific world, visiting places such as the Massachusetts Institute of Technology (MIT) in Cambridge, Cold Spring Harbor Laboratory and the local Gulbenkian Institute, which was funded by another Portuguese foundation established by a wealthy benefactor, Calouste Gulbenkian.

"I didn't know Champalimaud as a person, so I had to try to reconstruct in my mind what he might have wanted," Beleza says. Two things occurred to her immediately. Being an entrepreneur, Champalimaud would probably like to support translation of results from the bench to the clinic, she thought. And since he was nearly blind when he died at 86, it would be appropriate to support research into vision.

Beleza's tour refined her initial ideas. She recalls having an epiphany when talking with MIT's Nobel laureate Susumu Tonegawa, who told her that vision was too narrow an area for a research institute. "He explained how much our vision is controlled by our brains and that we need to understand much more basic neuroscience before we can start systematically translating results to the clinic." This and other conversations convinced Beleza

that a large part of the money should support basic research without any requirement for translation or, indeed, any strings attached at all.

Coutinho offered to help develop a competitive neuroscience programme for the centre, along with temporary lab space for new recruits, allowing them to start without delay. And he encouraged her to also consider cancer research, an area that is more readily translatable. It is also particularly appropriate in Portugal, where cancer care there is among the poorest in Europe, according to a 2005 consultancy report on health care commissioned by the Champalimaud Foundation.

In the final plan, Beleza settled on an annual prize of €1 million (US$1.6 million) for vision — to be awarded for research or for preventing blindness — together with the new research centre, which will combine clinical cancer research and basic neuroscience with a focus on identifying the neural circuits involved in driving behaviours.

The decision came as a disappointment to many scientists around Portugal who had lobbied for a more general distribution of the wealth. But Beleza was adamant about pursuing the big project. Her conversations with the scientific élite had convinced her of the value of focusing resources. They had also shown

Leonor Beleza was asked to head the Champalimaud Foundation.

Entrepreneur António Champalimaud left a half-billion-euro endowment for biomedical research.

her that good scientists will move to Lisbon only if the conditions on offer are better than those elsewhere. One advantage that the new principal investigators will have, for example, is a 'parachute fund'. This will allow them to take some of their allocated research money with them if they decide to bail out of the Champalimaud institute — an important provision for scientists who may be wary about joining a new and unproven institute.

Changing tack

This, and the opportunity to do blue-sky basic research, are tempting Portuguese expatriate Rui Costa, currently at the National Institutes of Health in Bethesda, Maryland, to move permanently to the Champalimaud Institute. Costa already has an adjunct position there, examining the dopaminergic circuits involved in movement. "It's nice that the Champalimaud people will allow you not only to do good science, but to do it under new conditions," he says. So far Beleza has recruited just three principal investigators and not, as yet, a director. She says she wants the best scientists from all over the world, but "it would be nice if some were Portuguese".

Reaction to the centre has generally been positive, although it is still early to say how well it will compete with established research institutes or whether the mix of cancer and neuroscience will prove fruitful. With no thought to modesty, top award-winning architect firm Charles Correa Associates — who built the Brain and Cognitive Neuroscience Complex at MIT — has been contracted to build a research institute with a cancer clinic worthy of the project's grandiose aims. The target completion date of 2010 will coincide with the centenary of Portugal's republic. Until then, on the foundation's webpage, the virtual fly-through of the architectural plans provides an exhilarating ride.

Outside, the complex boasts a broad public plaza that ascends gradually alongside the building to yield a sudden, stunning view of the Atlantic Ocean. The design is a fitting symbol of voyages into the unknown, of both fifteenth-century ocean explorers and twenty-first-century researchers. And Mainen, for one, is excited about the discoveries ahead. "I never expected to end up here," he says, "but I hope this project will plant Portugal's flag on the research map." ∎

Alison Abbott is *Nature*'s senior European correspondent.
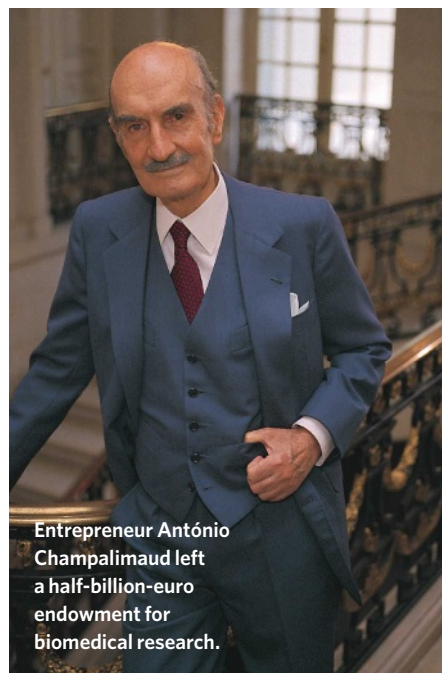
Radical rethink is needed on climate-change policy

SIR — The irreconcilable differences between David S. Reay's Book Review of *The Hot Topic* (*Nature* **452,** 31; 2008) and mine, expressed in *Nature Reports Climate Change* (see www.nature.com/climate/2008/0804/full/climate.2008.23.html), go to the heart of why there is now a crisis in climate policy. Reay seems to believe that agreement with a normative agenda precludes the need for rigorous evaluation of evidence or of proposed policy actions, and so falls into the same traps as Gabrielle Walker and David King, the authors whom he praises.

These authors have no doubt that the Kyoto Protocol is the road to follow. They consider that anyone, particularly an American, who doesn't agree is wrong — and perhaps even corrupt.

However, the Kyoto approach is broken, as my colleague Steve Rayner and I have pointed out (*Nature* **449,** 973–975; 2007). The 'bigger and better Kyoto' formula promoted by the authors — and by Al Gore, the European Union and the British Government — was defeated in Bali in December.

Instead, the geopolitical centre of gravity moved from Europe into the Pacific. There, four major powers will determine the future shape of global climate policy: China, India, Japan and the United States, supported by Canada and Australia (once that country recovers from the excitement of the symbolic act of 'signing Kyoto').

The framing concept for future climate policy must now shift, recognizing the failure of Kyoto to manage demand by imposing output targets. The reduction of energy intensity across all economies, including those of China and India, focusing first upon those primary sectors with the heaviest energy use, is emerging as the most robust replacement.

This approach not only preserves competitiveness but also has a demonstrated capacity to deliver results. It is now being promoted with increasing vigour outside Europe.

The reality is that we are at a tipping point for climate policy. Now is a moment where new ideas and new options arising from a radical rethink of the dismal record so far could have a dramatic and positive effect. Instead of welcoming such a discussion, the authors of *The Hot Topic* seem to wish to prevent it by continuing a tired political campaign for an approach that has not worked and cannot work.

Where climate change is concerned, it is not enough to identify good guys and bad guys — we must also be able to discern good policies from bad. The authors and reviewer of *The Hot Topic* conflate their judgements of the former with the latter.

Gwyn Prins
Mackinder Centre for the Study of Long Wave Events, London School of Economics, Houghton Street, London WC2A 2AE, UK

Comparing the legacies of Gauss, Pasteur and Darwin

SIR — Kevin Padian's enthusiastic Essay on Charles Darwin ('Darwin's enduring legacy' *Nature* **451,** 632–634; 2008) asks whether any single individual made so many lasting contributions to a broad area of science as Darwin did to biology. Let us remember that the nineteenth century also included Carl Friedrich Gauss, often called the greatest mathematician since antiquity, and Louis Pasteur, sometimes considered humanity's greatest benefactor because of his (and Robert Koch's) germ theory of disease.

It is a straightforward exercise to counter Padian's top ten darwinian topics (all of them evolution-oriented) with a much broader list for Gauss. He profoundly influenced modern life with his fundamental breakthroughs in statistics, algebra, analysis and other fields of mathematics — the 'queen of sciences'. His insights permeate all areas of science and engineering, including the theory of evolution.

Without Pasteur's revolution in medicine, many beacons of social and intellectual life would not have survived to formulate their thoughts. So, although Darwin was certainly one of the greatest, he had some even more influential contemporaries.

Jürgen Schmidhuber
IDSIA, Galleria 2, 6928 Manno, Switzerland, and Fakultät für Informatik, TU München, 85748 Garching bei München, Germany

Spanish funds will finance Bilbao's spallation bid

SIR — As the president of the executive commission of the proposed European Spallation Source (ESS) neutron-science facility in Bilbao, Spain, I appreciate the comments in your Editorial 'The big picture' (*Nature* **450,** 586; 2007) about European endeavours to construct and finance the ESS. Although the commission acknowledges that there is a tendency in other countries to use European funds to finance such scientific projects, this does not apply in our case.

The joint bid by Spain and Hungary does not propose the use of European funds for this purpose. Our intention is simply to work together to facilitate approval for the construction of this important project, in the absence of specially allocated European funds (see www.essbilbao.com).

As a candidate host for the ESS, Bilbao depends exclusively on funds provided by the Spanish central government and the Basque government. These funds will provide sufficient core financial support to implement the project with a clear guarantee.

Juan Urrutia
Executive Commission of ESS-Bilbao, Gran Vía 35, 48009 Bilbao, Spain

Give south Indian authors their true names

SIR — Your recent News Feature 'Identity crisis' (*Nature* **451,** 766–767; 2008) reveals that many Chinese, Japanese and Korean researchers do not receive due credit for their work, owing to inconsistent abbreviation practices and journal requirements regarding names. Many other Asian scientists face similar problems.

We are three female postgraduate students of south Indian ancestry. Indians from the south traditionally do not have surnames. It is only when forced to comply with Western naming standards that they use their father's given name as a substitute. As a consequence, journal rules require them to publish research under the fathers' given names (with which we — Nalini, Jeevananthinee and Sujatha — also sign this Correspondence letter). Obviously, as young south Indian scientists making a contribution to science, we would prefer to be identified with our first names and not by our fathers' given names.

India produces more than 100,000 postdoctoral scientists every year (see *Nature India* at www.nature.com/nindia/about/index.html). We believe that now is the time to introduce a consistent publication system that accommodates Indian names. The universal author-identification that uses contributor IDs, as discussed in your News Feature, is a good start. Such a system could be designed along the lines of the digital object identifier (doi) system used for journal articles. That could be followed by changes to reference rulings in journals to allow for citation of papers with single-name authors who are linked to a contributor ID.

We hope that all of science will take note of the extent of the Asian identity crisis in publishing and will work towards creating a universal system of authorship.
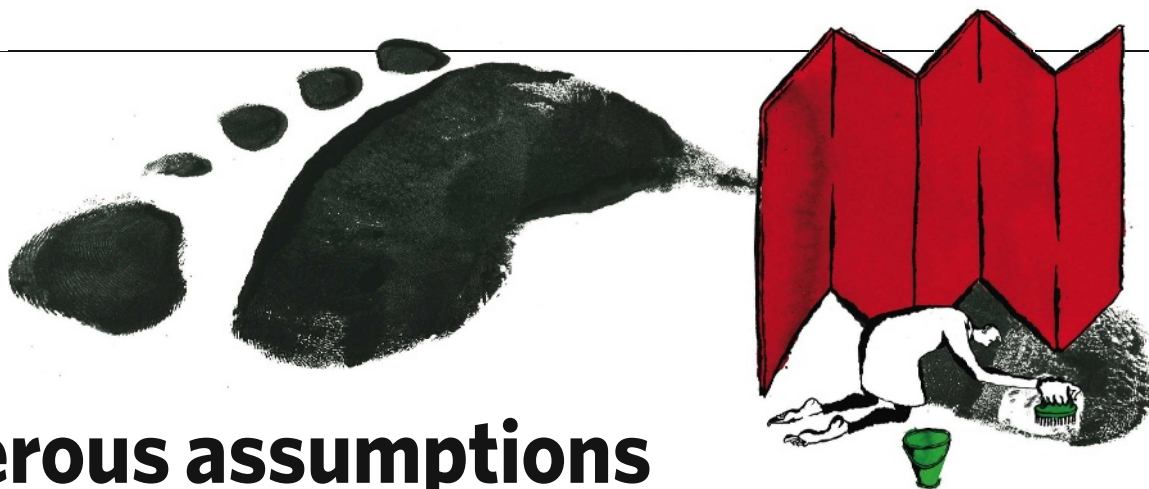
Nalini Puniamoorthy, Jeevananthinee Jeevanandam, Sujatha Narayanan Kutty
Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, Singapore

Contributions may be submitted to correspondence@nature.com. We also welcome comments and debate at Nautilus (http://blogs.nature.com/nautilus).

COMMENTARY

Dangerous assumptions

How big is the energy challenge of climate change? The technological advances needed to stabilize carbon-dioxide emissions may be greater than we think, argue **Roger Pielke Jr**, **Tom Wigley** and **Christopher Green**.

The United Nations Climate Conference in Bali in 2007 set the world on a two-year path to negotiate a successor to the 1997 Kyoto Protocol. Yet not even the most rosy-eyed delegate could fail to recognize that stabilizing atmospheric carbon-dioxide concentrations is an enormous undertaking. Here we address the magnitude of the technological changes required to meet that challenge. We argue that the size of this technology challenge has been seriously underestimated by the Intergovernmental Panel on Climate Change (IPCC), diverting attention from policies that could directly stimulate technological innovation.

The IPCC uses 'reference' scenarios of future emissions that assume no policy interventions directed towards reducing greenhouse-gas emissions (notably carbon dioxide) to determine the magnitude of additional emissions reductions ('mitigation') needed to stabilize atmospheric carbon-dioxide concentrations at various levels. It is on these additional reductions that policy-makers have focused most attention.

Here we show that two-thirds or more of all the energy efficiency improvements and decarbonization of energy supply required to stabilize greenhouse gases is already built into the IPCC reference scenarios. This is because the scenarios assume a certain amount of spontaneous technological change and related decarbonization. Thus, the IPCC implicitly assumes that the bulk of the challenge of reducing future emissions will occur in the absence of climate policies. We believe that these assumptions are optimistic at best and unachievable at worst, potentially seriously underestimating the scale of the technological challenge associated with stabilizing greenhouse-gas concentrations.

> **"The technological challenge has been seriously underestimated by the IPCC."**

The reference scenarios used by the IPCC's fourth assessment report (AR4) were described in a 2000 Special Report on Emission Scenarios[1] (SRES). In 2003, the IPCC decided not to develop comprehensive, new scenarios for AR4, so they used the SRES scenarios and related pre- and post-SRES scenarios[2,3] based on similar socioeconomic assumptions.

Climate scientists have argued that the outdatedness of the SRES scenarios is not that important when running them through climate models, because the scenarios cover a wide range of possible future emissions. But for IPCC Working Group III, which is concerned with mitigation of climate change, the details of emissions scenarios matter a great deal for considering policy options.

To assess the full challenge of reducing future emissions in line with particular stabilization targets, we begin with a 'frozen technology' baseline[4–6], which assumes that future energy needs are met with the technologies available in some baseline year (the technologies are 'frozen' in time). This approach differs from the SRES scenarios, which include various rates of spontaneous decarbonization[4,7].

The IPCC Working Group III links carbon-dioxide emissions to four specific 'drivers': population[1]; economic activity (gross domestic product or GDP) per capita; energy intensity (primary energy consumption per unit of GDP); and carbon intensity (carbon-dioxide emissions per unit of energy). These four elements are the building blocks for all emissions scenarios, and are widely used in climate-change assessments including efforts to estimate the costs of mitigation.

Decarbonization of the global energy system depends mainly on reductions in energy intensity and carbon intensity. These result from technological changes that improve energy efficiency and/or replace carbon-emitting systems with ones that have lower (or no) net emissions.

The true baseline

We also use the emissions-scenario building blocks in our analysis, but adopt a frozen-technology baseline to reveal the full challenge of decarbonization. Using this baseline also reveals the huge amount of emissions-reducing technological change built into the SRES and similar scenarios. Built-in emissions reductions were discussed briefly in AR4 by Working Group III (ref. 4), but are not reflected in its Summary for Policymakers or elsewhere. The significance of starting with a frozen-technology baseline is not yet widely appreciated.

Figure 1 (overleaf) shows the assumptions in the IPCC AR4 report for future emissions reductions during the twenty-first century, consistent with a carbon-dioxide stabilization target of about 500 parts per million. In the Working Group III report, the IPCC observes that "there is a significant technological change and diffusion of new and advanced technologies already assumed in the baselines"[4].

But how much is "significant"? The median of the reference scenarios considered by the IPCC AR4 (righthand bar, Fig. 1), requires 2,011 gigatonnes of carbon in cumulative emissions reductions to stabilize atmospheric carbon-dioxide concentrations at around 500 parts per million (the blue and red portions of the AR4 bar). This scenario also assumes that 77% of this reduction (the blue portion) occurs spontaneously, whereas the remaining 23% (the red portion) would require explicit policies focused on decarbonization.

These assumptions are robust across the scenarios used by the IPCC. Figure 1 also
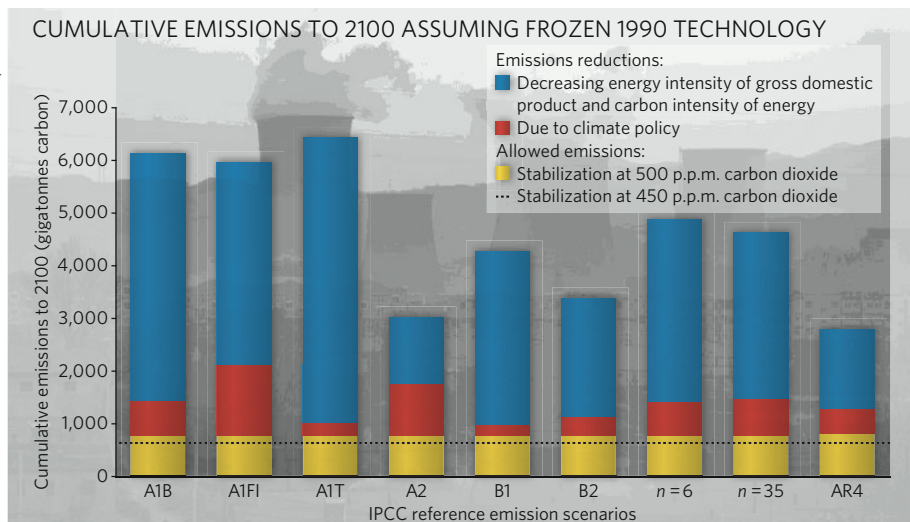
B. MELLOR

Figure 1 | **Cumulative emissions.** A range of 'built-in' emissions reductions (blue) in the scenarios used by the Intergovernmental Panel on Climate Change (IPCC). Total cumulative emissions to 2100 associated with a frozen-technology baseline are shown for: six individual scenarios, the means of these scenarios (n=6), and for all 35 IPCC scenarios, and the median of the scenario set (AR4). Additional reductions will have to be achieved by climate policy (red), assuming carbon-dioxide stabilization at about 500 parts per million (p.p.m.), leaving allowed emissions for this stabilization target (yellow).

shows the same breakdown in the 6 'illustrative' SRES scenarios[1] (plus the mean of these 6 and of the 35 complete SRES scenarios considered by the IPCC). In all cases, the IPCC assumes that most of the challenge (between 57% and 96%) of achieving stabilization at around 500 parts per million will occur automatically, leaving a much smaller emissions-reduction target for explicit climate policies.

Unpredictable future

The IPCC scenarios include a wide range of possibilities for the future evolution of energy and carbon intensities. Many of the scenarios are arguably unrealistic and some are likely to be unachievable. For instance, the IPCC assumptions for decarbonization in the short term (2000–2010) are already inconsistent with the recent evolution of the global economy (Fig. 2). All scenarios predict decreases in energy intensity, and in most cases carbon intensity, during 2000 to 2010. But in recent years, both global energy intensity and carbon intensity have risen, reversing the trend of previous decades.

Most SRES scenarios also predict a rapid decline in energy intensity (exceeding 1.0% per year), which may be neither realistic nor achievable. To achieve a century-long 1.0% annual rate of energy intensity decline requires very large increases in energy efficiency[8]. Even with a substantial policy effort this would be very difficult to achieve. Only about 20% (± 10%) of global energy intensity decline can be expected from sectoral shifts in economic activity, such as from manufacturing to services[8]. The rest must come from improved efficiencies in individual energy-using sectors, requiring either technology changes or new technologies.

One reason for the current rise in global energy and carbon intensities is the economic transformation taking place in the developing world,

especially in China and India. As development proceeds, rural populations move to high-rise buildings that consume energy and energy-intensive materials. This process is likely to continue, not only in these countries, but all over populous south Asia, and eventually Africa, until well beyond 2050. An analysis of China's carbon-dioxide emissions estimated them to be rising at a rate of between 11% and 13% per year[9] for the period 2000–2010, which is far higher than that assumed by the SRES scenarios for Asian emissions (2.6–4.8% per year).

Because of these dramatic changes in the global economy it is likely that we have only just begun to experience the surge in global energy use associated with ongoing rapid development. Such trends are in stark contrast to the optimism of the near-future IPCC projections
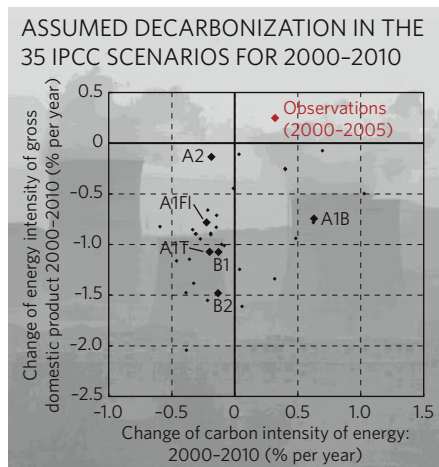

Figure 2 | **Decarbonization discrepancies.** Implied rates of carbon- and energy-intensity decline from the 2000 Special Report on Emission Scenarios, showing six illustrative scenarios. The red marker indicates actual observations (2000–2005) based on global economic growth calculated using market exchange rates.

and seem unlikely to alter course soon. The world is on a development and energy path that will bring with it a surge in carbon-dioxide emissions — a surge that can only end with a transformation of global energy systems. We believe such technological transformation will take many decades to complete, even if we start taking far more aggressive action on energy technology innovation today.

Enormous advances in energy technology will be needed to stabilize atmospheric carbon-dioxide concentrations at acceptable levels. If much of these advances occur spontaneously, as suggested by the scenarios used by the IPCC, then the challenge of stabilization might be less complicated and costly. However, if most decarbonization does not occur automatically, then the challenge to stabilization could in fact be much larger than presented by the IPCC[10,11].

The IPCC plans to update the SRES for its next report (due in 2013 or later), but in the meantime climate policy would be better informed by having a clear view of the size of the technological challenge.

There is no question about whether technological innovation is necessary — it is. The question is, to what degree should policy focus directly on motivating such innovation? The IPCC plays a risky game in assuming that spontaneous advances in technological innovation will carry most of the burden of achieving future emissions reductions, rather than focusing on creating the conditions for such innovations to occur. ∎

Roger Pielke Jr is in the Center for Science and Technology Policy Research, University of Colorado, Boulder, Colorado 80309-0488, USA. Tom Wigley is in the National Center for Atmospheric Research, Boulder, Colorado 80305, USA. Christopher Green is in the Department of Economics, McGill University, Montreal H3A 2T7, Canada.

1. *Special Report on Emissions Scenarios* (eds Nakicenovic, N. & Swart, R.) (Cambridge Univ. Press, Cambridge, UK, 2000).
2. Hanaoka, T. et al. Greenhouse Gas Emissions Scenarios Database and Regional Mitigation Analysis GCER Report (Natl Inst. Environ. Stud., Tsukuba, Japan, 2006).
3. Nakicenovic, N., Kolp, P., Riahi, K., Kainuma, M. & Hanaoka, T. *Environ. Econ. Pol. Stud.* **7,** 137–173 (2006).
4. Fisher, B. S. et al. in *Climate Change 2007: Mitigation, Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Metz, B., Davidson, O. R., Bosch, P. R., Dave, R. & Meyer, L. A.) Ch. 3, 218–220 (Cambridge Univ. Press, Cambridge, UK, 2007).
5. Edmonds, J. A., Wilson, T. & Rosensweig, R. *A Global Energy Technology Strategy Project Addressing Climate Change: Initial Findings from an International Public–Private Collaboration* (Joint Global Change Research Inst., College Park, Maryland, 2001).
6. Edmonds, J. A. & Smith, S. J. in *Avoiding Dangerous Climate Change* (eds Schellnhuber, H. J. et al.) 385–292 (Cambridge Univ. Press, Cambridge, UK, 2006).
7. Green, C. & Lightfoot, H. D. *Making climate stabilization easier than it will be: The report of IPCC WG III*, 6–13 (McGill University, C²GCR Quarterly, 2002-1, 2002).
8. Baksi, S. & Green, C. *Energy Pol.* **35,** 6457–6466 (2007).
9. Auffhammer, M. & Carson, R. T. *J. Environ. Econ. Manag.* (in the press)
10. Hoffert, M. I. et al. *Nature* **395,** 881–884 (1998).
11. Hoffert, M. I. et al. *Science* **298,** 981–987 (2002)

See Editorial, page 503 and News, page 508.
Join the discussion at: http://tinyurl.com/ys9xh8

BOOKS & ARTS

From spies to servers

Fastidious history chronicles how a suburb of Washington DC became home to the Internet.

Internet Alley: High Technology in Tysons Corner, 1945–2005
by Paul E. Ceruzzi
MIT Press: 2008. 192 pp. $29.95, £19.95

Joel Shurkin

For a brief, unhappy time, I worked in the American intelligence community in the suburbs west of Washington DC in northern Virginia. The area was littered with clusters of office buildings, many with no names on the side. It was, and still is perhaps, the largest accretion of mediocre architecture and spies on the planet.

Lifts went to floors that did not exist on the lobby directories. Many had unmarked windowless offices within offices, Sensitive Compartmentalized Information Facilities equipped with secure telephones, computers, shredders and copiers. Men and women scurried to and fro, tucking their identification badges in their shirt pockets when they left the buildings for lunch. Colours told how high your security clearance was. I was advised never to go to the same restaurant on a regular schedule lest I be followed. Followed by whom, I have no idea. Expensive cars filled the parking lots. I fantasized that in my next life I would be the Lexus dealer in Tyson's Corner.

The area, west of the Potomac river, is one of the fastest growing, best educated and richest in the United States It is a giant suburb that feeds mostly off the federal government. The Central Intelligence Agency headquarters is there in Langley, the National Reconnaissance Office is in Chantilly, and scores of military think-tanks, labs and large unmarked subsidiaries of the Department of Homeland Security are scattered around.

The region is also the home of the Internet, as Paul E. Ceruzzi, curator of the National Air and Space Museum at the Smithsonian Institution, Washington DC, emphasizes in the title of his new history of the area. How it got that way is the subject of *Internet Alley*. In a nutshell: gigantic quantities of government money.

Much of the growth, Ceruzzi explains, originated from the cold war and the government's attempt to move as many strategic targets out of range of the Soviet A-bombs that would surely target the US capital. The centre of this activity was Tysons Corner — once dairy farmland that real-estate speculators bought and subdivided. After the Second World War, developers convinced the government to move many of its

Non-descript blocks at Tysons Corner, northern Virginia, are the historic home of US high-tech.

operations research centres there, soon to be followed by other defence contractors. Pleasant suburbs, such as the planned community of Reston, housed the workers. The area was late to the party: the government had already built substantial science centres across the Potomac in Maryland, including facilities for the National Institutes of Health, the National Bureau of Standards and later NASA.

In 1964, came the Beltway, the highway that circles Washington DC, triggering commercial development at the exits. Dulles International Airport, opened in 1962, added to the mix. Technology companies doing government business set up shop on or near the airport's access road, putting them close to the capital, the Pentagon and the airport.

The Internet grew up almost incidentally. ARPANET, the Defense Department's inter-computer system, morphed into the National Science Foundation's network and finally the Internet. At one time in the 1990s, 70% of domestic Internet traffic went through the parking garage of a non-descript building in Tysons Corner. By 1998, five of the 13 root servers of the Internet were in the Washington DC area, three of them in the same office park in Herndon, including the server with the master directory. Northern Virginia also became the home to enterprises such as AOL and MCI because of their proximity to the Internet infrastructure.

In March 2000, the Internet bubble burst

and many commercial firms went belly up. The main economy of Internet Alley, meanwhile, suffered little. As Ceruzzi points out, most of the firms in northern Virginia, unlike those in Silicon Valley, were not famous for risk or innovation and had only one customer: the government, which is generally not affected by such market vicissitudes. Ceruzzi chronicles the evolution of Internet Alley astutely and accurately, if sometimes too fastidiously. Weird then that such a book has no index (a grave offence), even if, happily, it is illustrated.

When 9/11 struck, the federal spigot became a power hose. Much of the spending is in the 'black budget', the classified part of the federal budget. But the cars in the parking lot are fancier than ever and Tysons Corner has one of the most up-scale shopping centres in the United States, home to the first Apple retail store.

Yet, the whole place could disappear, Ceruzzi reveals. As a commercial concentration in the suburbs — what urban planners call an 'edge city' — Tysons Corner may be doomed. You can't walk anywhere in an edge city, so if you want urban amenities, you move elsewhere. People who want peace and quiet are decamping farther out and away from Tysons Corner. And the commute is terrible. ■

Joel Shurkin is Sneddeden Chair in Journalism at the University of Alaska Fairbanks, Fairbanks, Alaska 99775. He is author of *Broken Genius: the Rise and Fall of William Shockley*.

THEATRE
Ten years of science off Broadway

Alan Packer

Breaking down the walls that divide scientists from the rest of society is the goal of the First Light Festival of plays about science and technology, produced by the Ensemble Studio Theatre in New York City, with the support of the Alfred P. Sloan Foundation. The festival celebrates its tenth anniversary this month.

Each year, the theatre solicits proposals for science-related plays, especially those that challenge stereotypes. "The ones that get lost in the research tend to be dry, but those that portray scientists as fully three-dimensional human beings, and show the context of their lives, are more successful," says current artistic director William Carden.

A dozen projects are awarded grants ranging from US$500 to $10,000, and the annual festival presents the plays in various stages of production during its month-long run. Playwrights are paired with academic advisers, such as physicist Brian Greene and biologist Darcy Kelley.

This year's productions include *Pure* by Rey Pamatmat, about the British mathematician Alan Turing; *The Flower Hunter*, a specially

Damien Atkins's 2007 play *Lucy* explores teenage autism.

commissioned work by distinguished playwright Romulus Linney, which dramatizes the life and work of William Bartram, one of America's first naturalists; and Amy Fox's play about sudden infant death syndrome, *By Proxy*.

Founded as a developmental theatre in 1972 by the late Curt Dempster, who served as its artistic director for 35 years, the Ensemble Studio Theatre's reach is much greater than its modest home on West 52nd Street in Manhattan suggests. It nurtures new artists and plays by commissioning works and offering ongoing development and a venue.

Past plays have featured celebrated actors such as Cynthia Nixon, and have gone on to theatres around the world to much acclaim. Noteworthy productions include Arthur Giron's *Moving Bodies*, which dramatized physicist Richard Feynman's colourful life and career, Paul Mullin's *Louis Slotin Sonata*, about the fatal slip of a Los Alamos physicist, Bob Clyman's *Secret Order*, on battles between scientists on the verge of a breakthrough in cancer research, and *String Fever* by Jacqueline Reingold, which infuses string theory into the romantic life of the main character.

A *New York Times* reviewer wrote in 2006 that the series was "always guaranteed to send audiences out with plenty to think about". Although the cultural impact is difficult to gauge, exposing theatregoers to nuanced portrayals could reshape the view of scientists in the popular imagination. ∎

Alan Packer is senior editor at *Nature Genetics.*
The First Light Festival runs until 26 April (www.ensemblestudiotheatre.org).

HISTORY
An astrolabe for the people

Philip Ball

The fate of a rare fourteenth-century astrolabe quadrant — a kind of medieval pocket calculator — hangs in the balance while the British Museum in London tries to raise £350,000 (US$700,000) to acquire it before its deferred release to a private buyer. The palm-sized brass device can be used to determine the time of day from the position of the Sun and to calculate the date of Easter, among other functions.

The simplicity of the instrument suggests that, in the Middle Ages, some sections of English society were surprisingly literate in basic mathematics and astronomy. Most surviving astrolabes, whether standard disc-shaped devices or quarter-circle quadrants, are elaborate instruments capable of astrological calculations that require specialist academic knowledge. This one is a stripped-down, everyday item — something that a cleric or a merchant could have carried for convenient time-keeping. "You had to know some

Palm-sized 'calculator' for medieval pockets.

astronomy to work one of these devices," says Jim Bennett, director of the Museum of the History of Science in Oxford, UK. It suggests, he says, that people of the time "had a closer astronomical awareness than we do now".

Found in 2005 at Canterbury, UK, the quadrant has been dated to about 1388, around the time that Geoffrey Chaucer began to write *The Canterbury Tales*. Chaucer was highly informed about astronomy and astrology. In 1391, he wrote a treatise on astrolabes that became the standard reference text for several centuries. The Canterbury quadrant was excavated from beneath a series of clay floors on the site of an old inn, just outside the city's walls on the main road towards London. Perhaps it was lost at the inn by a merchant travelling to or from Canterbury, rather like Chaucer's pilgrims.

The instrument was initially put up for sale in 2007 by the London auctioneers Bonhams, and was expected to fetch £60,000–£100,000. Subsequent private

dealings led to an agreed sale at more than three times the original figure. The identities of the dealer and the latest buyer have not been made public.

Because of the cultural importance of the astrolabe quadrant, Britain's culture minister, Margaret Hodge, was persuaded by a review committee to defer granting an export licence for it until June 2008 to give the British Museum time to try to buy the instrument for its new medieval gallery. Such deferrals are more usually applied to works of fine art than to scientific items.

The existence of this practical device in Chaucer's time sheds new light on his astrolabe treatise. His dedication of the book to his son had left scholars bemused. It seemed hard to believe that a young person would have understood how to use an astrolabe. The Canterbury quadrant "supports the idea that Chaucer could write such a treatise at a popular level", Bennett explains. "It suggests that this kind of knowledge wasn't too arcane or academic." ∎

Philip Ball is a consultant editor at *Nature*.

M. MARGOLIN

Meryl Streep plays the patron of physics student Liu Xing in *Dark Matter*.

FILM

Celluloid explores life of graduate student killer

Emma Marris

In 1991, Gang Lu, a Chinese graduate student studying space plasma physics at the University of Iowa, shot and killed five people, wounded one, and then shot and killed himself. Three of the victims were on his dissertation committee; letters he left indicated that he shot them because they did not think highly enough of his dissertation.

The crime deeply affected Chen Shi-Zheng, then a graduate student in the performing arts at New York University. "The event shaped me profoundly for years," he says. Chen identified with the troubled gunman who, like him, had come from China to study in the United States. "Foreign students are just passengers," he says. "You are not human beings; you get overlooked completely."

Now a successful opera director, Chen has returned to the crime that haunted him with the film *Dark Matter*. Released this month in the United States, it is a fictional treatment of the Gang Lu story — one that humanizes the shooter to an uncomfortable degree.

Rather than exploring what makes someone commit such a terrible crime, the film invites the world to comprehend the pressured lives of foreign students. "The tragic ending is the point of departure that gets people thinking about graduate student life," says Chen. "I want people to be as shaken as I was." But, he insists, he did not intend to make a campus-violence film. "It was intended to be about the Chinese experience."

Many Chinese graduate students have high expectations about how their lives should pan out. It rarely goes quite as planned, explains Chen. They may work long hours and feel socially isolated. To capture this invisibility metaphorically, Chen switched the student's field to astrophysics, specifically dark matter, the unseen substance prevalent throughout the Universe.

Chen and the screenwriter worked with astrophysicist David Weinberg from Ohio State University in Columbus to get the science right. Weinberg helped to stage a plausible academic disagreement between arrogant adviser Jacob Reiser (played by Aidan Quinn) and brilliant student Liu Xing (played by Liu Ye).

"My notion," says Weinberg, "was that Reiser had become famous by coming up with a creative idea for the initial conditions for the early Universe, but the data were turning against him and he was in denial about that."

Xing goes off on his own, dropping the 'Reiser model' and becoming more interested in the interplay of superstring theory and the identity of the dark matter particle. Reiser, selfishly, is not pleased.

Xing is portrayed as a sympathetic, even charming, character; a victim very much wronged by his department. During most of the film he hangs out with his friends who toil at their theories and unwind with a little table football, beer and idle chatter about women. Meryl Streep plays Joanna Silver, a wealthy patron of the university and Xing's confidante.

The film does such a good job of making its protagonist lovable that it risks seeming to defend or even glorify the horrible murders that Lu committed. The counterparts to Lu's real victims are here played as villains.

Will cosmologists walk out of the film enraged by scientific inaccuracies? Weinberg says no. "If you are going to be infuriated, it will be over the question of whether it is justifying the events."

"Everyone says they didn't see it coming,'" says Chen of early reaction to the film's violent end. "In the real case, no one saw it coming." ∎
Emma Marris is a correspondent for *Nature*.

***Dark Matter* opens in the United States on 11 April (www.darkmatterthefilm.com).**

NEWS & VIEWS

GENOMICS

When the smoke clears …

Stephen J. Chanock and David J. Hunter

Three studies identify an association between genetic variation at a location on chromosome 15 and risk of lung cancer. But they disagree on whether the link is direct or mediated through nicotine dependence.

The characterization of variation in the human genome has begun to pay dividends. These are mainly being delivered through genome-wide association studies (GWAS), provided that they are sufficiently large and are followed by replication of the notable findings. Since early 2007, variations at nearly 100 regions of the genome have been associated with an increased risk for diseases with a complex genetic background, such as diabetes, inflammatory bowel disease, cancer (most notably breast, colorectal and prostate) and heart disease.

The latest instalment in the GWAS story deals with lung cancer, and comes in the form of three papers — two published in this issue[1,2], on pages 633 and 638, and one in *Nature Genetics*[3]. All of them identify variation in the same region of the long arm of chromosome 15 (15q24/15q25.1) as the 'top hit' for genomic association with lung cancer. Among the genes in this region are those that encode subunits of nicotinic acetylcholine receptors, which as their name suggests have an affinity for nicotine. The genetic variation analysed was in the form of single nucleotide polymorphisms (SNPs). These are DNA sequence variations that arise from the substitution of one nucleotide base for another, and overall constitute approximately 90% of common variation in the human genome.

The three studies are all large and appropriately replicated, and provide strong evidence for an association between SNP variation at 15q24/15q25.1 and lung cancer. But they differ on whether the connection is direct or mediated via smoking behaviour — that is, characteristics such as the duration and 'dose' of lifetime smoking, and/or the propensity for nicotine addiction. Previous studies[4,5] had identified the genes encoding subunits of nicotinic acetylcholine receptors as strongly associated with smoking behaviour. The association between smoking and lung cancer is among the strongest in the epidemiological firmament, and any gene variant that is modestly linked with smoking behaviour will seem to be associated with lung cancer unless the matching of cancer cases and controls by smoking behaviour is close to perfect.

Thorgeirsson *et al.*[2] report an association of SNP variation at 15q24/15q25.1 with the number of cigarettes smoked per day and a nicotine-dependence scale. They suggest that the link with lung cancer is primarily mediated through nicotine dependence and thus provides a case study of a gene–environment correlation in the pathogenesis of disease. The two other groups, Hung *et al.*[1] and Amos *et al.*[3], examined SNP variation in patients with lung cancer and control subjects, and reach the opposite conclusion — namely, that the association is primarily with lung cancer and not with smoking.

In these kinds of studies, researchers distinguish between 'ever-smokers' and 'never-smokers'. Ever-smokers are current or former smokers; never-smokers are usually defined as those who have smoked fewer than 100 cigarettes in a lifetime.

Amos and colleagues[3] studied both cancer patients and controls mainly in ever-smokers. They describe the evidence for the association of their 'lead' SNPs with smoking behaviour (among ever-smokers) as "weak". Hung *et al.*[1] did not match cases and controls by smoking, but carefully examined the relation between the lead SNPs and smoking behaviour. In contrast to Thorgeirsson *et al.*, Hung *et al.* did not observe an association between their lead SNP and any of the measures of nicotine addiction, and thus suggest that the genetic association in their study is unlikely to be due to confounding effects of smoking behaviour.

The two studies[1,3] that also examined risk of lung cancer in never-smokers reached different conclusions, but we should consider this discrepancy preliminary. Hung *et al.*[1], with 352 cases among never-smokers, conclude that there is an association with the risk SNPs; Amos *et al.*[3] conclude that there is no such association, although with only 125 never-smoking cases in their series, statistical power was even more limited.

The extent of the disagreement in the conclusions of these studies is remarkable. Careful examination of the relation between the SNP variants in even larger data sets, with detailed measures of both smoking patterns and addictive behaviour, is needed to discern whether these discoveries relate to the risk of smoking, the risk of lung cancer or the risk of both. These are the first GWAS to attempt to identify the genetic component of a disease that has such an overwhelmingly strong environmental cause. They signal the need for greater methodological rigour in attempts to account for both the genetic and the environmental causes that we think underlie most diseases.

> **"These are the first genome-wide association studies to attempt to identify the genetic component of a disease that has such an overwhelmingly strong environmental cause."**

As GWAS have become available there has been something of a 'flight to quantity' as the importance of large sample sizes to detect modest effects has become apparent. Such sample sizes have often been obtained at the expense of collecting even minimal environmental and lifestyle information. Much GWAS research has tended to be retrospective, and thus this information could be biased; the requirements for public access to individuals' data may be suppressing the availability of potentially high-quality data sets because of limitations in obtaining informed consent. These studies[1–3] demonstrate the importance of balancing the undeniable need for quantity in GWAS with the need for quality of ancillary data, especially if we are to disentangle and understand the interplay between genes and environment.

The next round of research will involve resequencing 15q24/15q25.1 and incorporating any additional SNP variants into future epidemiological work. Moreover, further genome-wide scans are needed — especially in studies that can look at patients with lung cancer for whom detailed data on smoking exposure, nicotine dependence and duration and intensity of smoking are available. It is also essential to examine larger samples of non-smoking lung cancer patients and to perform analyses according to the specific type of cancer. Pooling of GWAS data will help, as will full disclosure of SNP rankings for the major smoking behaviours to permit a better understanding of the genetic associations with these behaviours (rankings based on

grouped data can be made fully available and disclosure poses no risk of loss of confidentiality for individual subjects).

Can findings such as these be applied in the clinic? Companies are already offering direct-to-consumer genetic testing for the risk of developing various diseases, using the same SNP chips as in the new studies[1–3]. The rationale claimed for determining risk in individuals is that it will encourage them to change their lifestyle and/or undergo screening so that disease can be detected earlier. In the case of smoking, however, there is a danger that such 'personalized' medical advice will weaken the public-health message that everyone should avoid smoking. Even if a subset of people are deemed 'resistant' to the effects of smoking on lung cancer development, it is unlikely that they will also be protected against the adverse effects of heart disease and obstructive pulmonary disease, disorders that are also associated with smoking. On the other hand, we may be able to evaluate smoking-cessation treatments informed by knowledge of a person's genetic predisposition to start smoking or to nicotine addiction, and thus add new weapons to the anti-smoking arsenal.

Follow-up studies should clear the smoke clouding the differing conclusions of these papers[1–3] and establish the biological rationale for the robust association of 15q24/15q25.1 with lung cancer. It is at least reassuring that all three groups point to the same region on chromosome 15. The sceptics have fretted that association studies would be riddled with false positive results; yet, because of the high standards that have been developed[6], the evidence for the association between some genetic loci and certain complex diseases is now unequivocal. For most diseases, more loci will surely be discovered, adding to the first wave of results that have been primarily related to disease causation. On the horizon we can see the crest of studies reporting on disease outcomes, but progress, both in understanding the basic causes and in estimating personal risks, will require environmental and lifestyle factors to be taken into account. To quote Winston Churchill, it is "perhaps the end of the beginning" in the battle to understand complex diseases. ∎

Stephen J. Chanock is in the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-4605, USA. David J. Hunter is in the Department of Epidemiology, Harvard School of Public Health, and the Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.
e-mails: chanocks@mail.nih.gov; dhunter@hsph.harvard.edu

1. Hung, R. J. et al. Nature 452, 633–637 (2008).
2. Thorgeirsson, T. E. et al. Nature 452, 638–641 (2008).
3. Amos, C. I. et al. Nature Genet. doi:10.1038/ng.109 (2008).
4. Saccone, S. F. et al. Hum. Mol. Genet. 16, 36–49 (2007).
5. Berrettini, W. et al. Mol. Psych. 13, 368–373 (2008).
6. Chanock, S. J. et al. Nature 447, 655–660 (2007).

EXTRASOLAR PLANETS

With a coarse-tooth comb

Gordon Walker

The search for Earth-like planets outside our Solar System is bedevilled by the lack of an adequate frequency standard for calibrating starlight. Tweaking existing laser 'frequency combs' could be a way forward.

As we look for planets orbiting other suns, it is often tiny, periodic shifts in the spectrum of light coming from a star — a tell-tale 'Doppler wobble' — that reveals the presence of one or more smaller, unseen companions. But when it comes to finding Earth-like planets, this technique reaches a hurdle: the lack of a suitable frequency standard with which to measure the truly tiny spectral shifts caused by such very small planets. On page 610 of this issue, Li et al.[1] take this matter in hand, adapting the optical comb — a technology developed to measure very high frequencies — to the purposes of astronomical measurements.

The Austrian physicist and mathematician Christian Doppler first suggested in 1842 that the difference in colour observed between the two stars in some binary systems could be explained if the stars were moving in opposite directions along our line of sight. The wavelength of the light from the star moving towards us would be shifted to shorter, bluer wavelengths, and that of the star moving away from us to longer, redder wavelengths.

In this case, Doppler was wrong: such a shift is much too small for the naked eye to detect, and the different colours of binary stars actually correspond to different surface temperatures. But he was right in principle. We are well attuned to the effect now named after him when it involves sound waves from bodies moving towards and away from us: the speed of sound is a million times slower than that of light, allowing our senses to pick up the difference in their tones. And spectrographs allow us to measure the relative displacement of the spectral lines of binary stars (Fig. 1a). The difference between the stars' velocities can easily be calculated from this wavelength shift, and plotting how this relative velocity changes with time gives the period of revolution of the stars and, with a little additional information, their relative masses and the size of their mutual orbit.

When the Doppler technique is applied to the search for extrasolar planets, only a single spectrum is measured — that of the putative parent star. If an orbiting planet is present, its gravitational attraction will at times pull the star towards us, and at times pull it away from us. The result will be a slow, periodic oscillation in the star's velocity relative to us, and an accompanying shift in its spectrum. The effect is tiny: Jupiter, the giant of our Solar System, has only one-thousandth of the mass of the Sun, and its gravity changes the Sun's radial velocity by just ± 13 m s^{-1} in a cycle that takes almost 12 years, the period of Jupiter's orbit. Earth, which is closer to the Sun, but 300 times smaller than Jupiter, affects the Sun's velocity to the tune of just ± 10 cm s^{-1} over the course of a year.

Planet-searching astronomers need to measure, with confidence and over timescales of years, the truly minuscule wavelength (equivalently, frequency) shifts that accord with
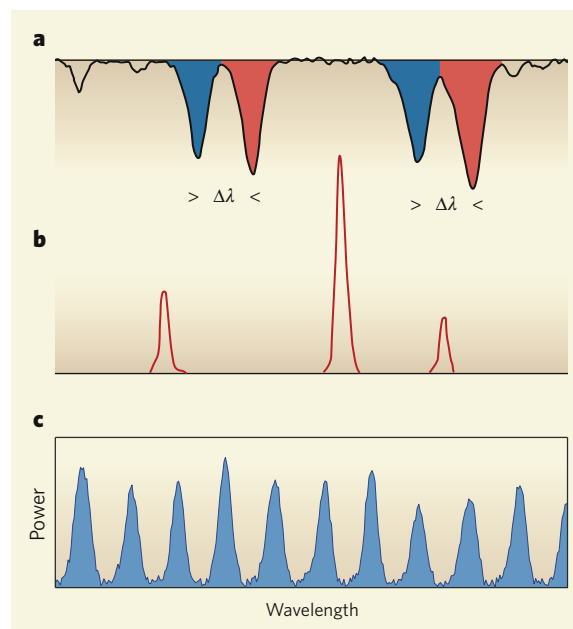

Figure 1 | Shifting stars. a, The motion of two stars in a binary system along our line of sight will shift the frequency of light observed by us: to longer, redder wavelengths if a star is moving away from us; to shorter, bluer wavelengths if the motion is towards us. **b,** Comparison with accurately calibrated reference lines (here from an iron arc) allows the degree of this wavelength shift, $\Delta\lambda$, to be measured. A similar measurement of the 'Doppler wobble' of a single star, induced by an unseen companion, is used in the search for extrasolar planets. **c,** Li et al.[1] provide a new reference standard based on a 'comb' of evenly spaced frequency lines that should allow these measurements to be made more accurately — perhaps opening up the possibility of spotting an Earth-like planet.

these velocity changes. So they need highly reliable local 'standards of rest' — standards of precisely known and completely stable wavelengths. Until now, discharge lamps and absorption spectra have served well. The first technique provides stabilized, reasonably well-distributed emission lines for elements such as thorium and argon (Fig. 1b); the second works by passing starlight through a cell containing a molecular vapour, usually iodine, so that the molecule's absorption lines are imposed on it. With further improvements such as the use of a highly stable, preferably evacuated, spectrograph, Doppler measurement precisions of about 1 m s^{-1} can be achieved using reference standards derived from either or both of these techniques.

The harvest from the radial-velocity method has been, at the time of writing, a crop of 261 giant planets. Some of these are rocky, and 25 of the solar systems involved have been found to have more than one planet[2]. The push is now on to find analogues of Earth — rocky planets within the habitable zone (the range of distances within which liquid water would be present) of their parent stars. The focus will be on stars that are both cooler and less massive than the Sun: in these solar systems the radius of the habitable zone is considerably smaller and closer to the star than it is in our own, and a planet of lower mass causes a larger velocity perturbation of the star over a shorter time.

Such cool stars are considerably brighter at infrared wavelengths than they are at optical wavelengths, and the intention is to build spectrographs specifically designed to cover this region. But therein lies a problem: at infrared wavelengths, there are no suitable extended emission or absorption spectra for use as a reference standard.

Enter the 'astro-comb'. Li and colleagues' brainchild[1] is an optically filtered comb of evenly spaced frequency references, all derived from a single frequency source — a pulsed laser. The idea is not new[3,4] (half of a Nobel Prize in Physics was awarded for the idea in 2005)[5], and nor is its application to astronomy[6]. But Li et al. are the first to realize the concept in a way suitable for astronomical practice, in what could be a breakthrough in the precision of astronomical spectroscopy.

The principal difficulty that the authors had to overcome was that, although the repetition rate of a fast pulsed laser generates a wide, stable comb of equally spaced sharp fringes, this spacing is much too fine to be resolved by the planet-search spectrographs currently in use or proposed. Li et al. use a Fabry–Perot optical filter to give a much sparser comb with teeth ideally spaced for optimum wavelength calibration (Fig. 1c). This might sound simple, but to cover the possible range of frequencies to be measured an effective astro-comb needs to provide at least a thousand teeth, each acting as a reference standard for a particular part of the spectrum, and each bright enough to be readily exploited in a short space of time. That

makes great technical demands on the available equipment.

In theory, the astro-comb could lead to radial-velocity measurements with a precision of 1 cm s^{-1} — nearly a hundred times better than the current best, and comparable to the magnitude of Earth's influence on the Sun. But the advance is only a first, albeit essential, step when it comes to the detection of terrestrial planets. There are many factors that can confound accurate measurements of the true radial velocity of a star: if our own Sun serves as a benchmark, a star's surface is a maze of bright granules caused by hot gas boiling to the surface at velocities of around 1 km s^{-1}; dark sunspots, coupled with the star's rotation, severely distort the measured radial velocity;

and rapid pulsations, a characteristic of most stars, further complicate matters. Extracting any signal corresponding to another Earth has substantial hurdles to overcome yet. ∎

Gordon Walker is emeritus professor in the Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, British Columbia V6T 1Z1, Canada.
e-mail: gordonwa@uvic.ca

1. Li, C.-H. et al. Nature 452, 610–612 (2008).
2. http://exoplanet.eu
3. Cundiff, S., Ye, J. & Hall, J. Sci. Am. 298 (4), 74–81 (2008).
4. Udem, Th., Holzwarth, R. & Hänsch, T. W. Nature 416, 233–237 (2002).
5. http://nobelprize.org/nobel_prizes/physics/laureates/2005/index.html
6. Murphy, M. T. et al. Mon. Not. R. Astron. Soc. 380, 839–847 (2007).

GEOPHYSICS

Humming a different tune

Toshiro Tanimoto

Earth breathes in and out, murmuring gently to itself as it does so. The habit has been ascribed to the tickling effects of ocean waves — but a new-found twisting oscillation might reopen the search for the source.

Over the past decade, the word 'hum' has acquired a special meaning for seismologists. No longer just what they might do under the shower, it connotes for them a fundamental resonant oscillation of the Earth. A sequence of these oscillation modes, with periods of between around 2 and 5 minutes, was first identified[1-3] in 1998. These were all 'spheroidal' modes, representing perturbations of the planet's equilibrium surface, rather akin to the effect of waves on water. Writing in *Geophysical Research Letters*[4], Kurrle and Widmer-Schnidrig now introduce a further, entirely different mode — 'toroidal' hum, in which parts of Earth's surface twist around in the horizontal plane (Fig. 1, overleaf).

The existence of this low-frequency Earth hum is not the surprising thing. Seismic noise is ubiquitous, generated by various natural processes such as falling water (the impact of, say, the Niagara Falls is not confined to the surface) and even swaying trees, as well as all manner of human activities. It is the magnitude of the hum that is disconcerting[5,6]: its summed amplitude is equivalent to a continuous earthquake of magnitude 6. (Because the waves are at such a low frequency, we humans cannot sense them; as they represent no threat to our well-being, there has presumably never been a need to evolve such a capability.) An earthquake of this size occurs once every three days on average; clearly, seismic activity cannot sustain hum of such magnitude and continuity.

Since those first intriguing findings, the ocean has by general consensus been identified as the most likely source of Earth hum:

the origin of the excitations seems to lie in oceanic areas at mid-latitudes, between about 30° and 60° north and south[7,8]. In addition, the amplitude of the effect has a periodicity of six months, with a maximum occurring in each hemisphere during its winter; satellite data show that ocean waves are particularly large at mid-latitudes during the winter months.

The proposal[9,10], which borrows an idea of some 60 years ago[11], is that so-called infra-gravity waves, which are known to have the same sort of periods as the hum[8,12], transmit this oceanic motion to the solid Earth. These waves are similar to tsunami waves — low-frequency, long-wavelength ocean waves that move the whole column of ocean waters, from surface to sea floor, as they propagate. The collision of such waves could produce large pressure variations[11], and thus excite the hum. A problem is that infragravity waves are mainly known to be a phenomenon of shallow water, although a mechanism for generating them in the deep ocean has recently been proposed[13].

Even so, a direct interaction between the atmosphere and the solid Earth has not been ruled out as a source of the hum. Atmospheric and oceanic effects are difficult to separate: when we see large-amplitude ocean waves, the cause is likely to be an atmospheric effect, namely strong winds. The observant frequent flyer from New York to Paris or Tokyo to San Francisco will note that, during winter in the Northern Hemisphere, flights are often diverted from the shortest geographical route, a great circle over the Arctic, to a more southerly route of near-constant mid-latitude. The
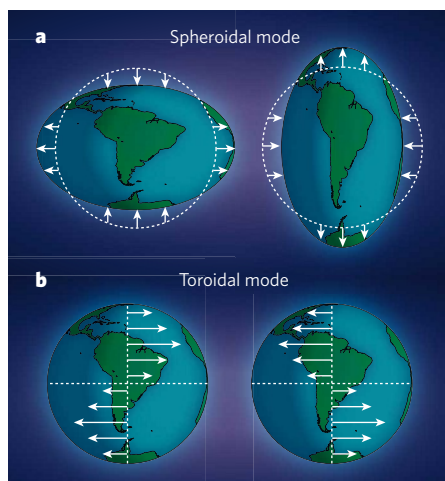
539

Figure 1 | A new twist. a, The first modes of Earth hum were discovered 10 years ago[1–3], and are spheroidal oscillations of Earth's surface, representing a perturbation in the vertical plane. These oscillations have been explained as the effect of the broadly analogous movements of ocean waves. **b,** Kurrle and Widmer-Schnidrig's discovery[4] of toroidal oscillation modes, in which different areas of Earth's surface twist in different directions, is less easily explained in the same way.

reason is the saving of one to two hours in flight-time thanks to strong westerly tail winds over the northern Pacific and Atlantic. On those occasions, the same watchful traveller might also, on looking out of the plane window, see rampant ocean waves far below.

Kurrle and Widmer-Schnidrig's analysis of Earth hum, with their discovery of toroidal modes[4], brings a new angle to these considerations of oceanic and atmospheric effects. All past work on hum has focused on spheroidal modes: measuring these modes is simpler, because one needs just a single instrument that measures seismic activity in the vertical plane. Twisting toroidal modes, on the other hand, require the analysis of seismograms in the two horizontal dimensions of Earth's surface. The interpretation of these seismograms is further complicated by the coexistence of spheroidal and toroidal modes in them, as well as noise generated by the local tilt of geological strata just under Earth's surface.

But the authors now show that there are peaks in the oscillation-frequency spectrum that correspond exactly to predicted frequencies of toroidal modes[14]. Whereas it is easy to see how the broadly analogous up-and-down action of ocean waves might produce spheroidal oscillation modes, it is less easy to see how oceanic infragravity waves might generate toroidal modes.

So what are the alternative explanations? Atmospheric theories developed to explain hum excitation in the past 10 years[6,15] have considered only the role of local variations in atmospheric pressure, which impinges vertically downwards at each point on Earth's surface. Again, this vertical force might help to explain the spheroidal modes, but it is

irrelevant to the excitation of toroidal modes. Fresh thinking is thus required, whether the source for the newly discovered modes lies in the atmosphere or in the oceans. One might speculate on possible mechanisms: perhaps winds exert shearing forces on the solid Earth through topographic coupling — when an air mass hits a mountain range, for instance — or perhaps long-period ocean waves hitting the undersea walls at continental shelves are generating horizontal forces.

Seismic noise is indeed ubiquitous, but, as Kurrle and Widmer-Schnidrig show[4], it also crops up in different forms. By learning more about the hierarchy of mechanical forces that act among the atmosphere, the oceans and the solid Earth, we might hope to become wiser about the origin of Earth's fundamental hum. ∎

Toshiro Tanimoto is in the Department of Earth Science, University of California,

Santa Barbara, California 93106, USA.
e-mail: toshiro@geol.ucsb.edu

1. Suda, N., Nawa, K. & Fukao, Y. *Science* **279,** 2089–2091 (1998).
2. Kobayashi, N. & Nishida, K. *Nature* **395,** 357–360 (1998).
3. Tanimoto, T., Um, J., Nishida, K. & Kobayashi, N. *Geophys. Res. Lett.* **25,** 1553–1556 (1998).
4. Kurrle, D. & Widmer-Schnidrig, R. *Geophys. Res. Lett.* **35,** doi:10.1029/2007GL033125 (2008).
5. Ekström, G. *J. Geophys. Res.* **106,** 26483–26493 (2001).
6. Tanimoto, T. & Um, J. *J. Geophys. Res.* **104,** 28723–28740 (1999).
7. Nishida, K. & Fukao, Y. *J. Geophys. Res.* **112,** doi:10.1029/2006JB004720 (2007).
8. Rhie, J. & Romanowitz, B. *Nature* **431,** 552–556 (2004).
9. Webb, S. *Nature* **445,** 754–756 (2007).
10. Tanimoto, T. *Geophys. J. Int.* **168,** 571–582 (2007).
11. Longuet-Higgins, M. S. *Phil. Trans. R. Soc. Lond. A* **243,** 1–35 (1950).
12. Tanimoto, T. *Geophys. J. Int.* **160,** 276–288 (2005).
13. Uchiyama, Y. & McWilliams, J. C. *J. Geophys. Res. Oceans* (in the press).
14. Gilbert, F. & Dziewonski, A. M. *Phil. Trans. R. Soc. Lond. A* **278,** 187–269 (1975).
15. Fukao, Y., Nishida, K., Suda, N., Nawa, K. & Kobayashi, N. *J. Geophys. Res.* **107,** doi:10.1029/2001JB000153 (2002).

DRUG RESISTANCE

The fight against fungi

André Goffeau

Although several drugs are available to combat often-deadly fungal infections, many of these pathogens have acquired multidrug resistance. Discerning how they have achieved this could help us hit back.

When we think of infectious agents, it is viruses and bacteria that first come to mind — fungi do not sound especially threatening. But opportunistic fungal infection is an emerging problem, particularly for immuno-compromised patients such as those with AIDS and cancer, or those who have undergone organ transplants. Fortunately, a number of effective antifungal drugs are on hand. But, like many other infectious microorganisms, pathogenic fungi have not taken this challenge silently, and have become resistant to not just one but several of these drugs. On page 604 of this issue, Thakur *et al.*[1] shed light on the molecular mechanism behind antifungal-drug resistance.

It has been known since 1974 that the yeast *Saccharomyces cerevisiae* can acquire resistance to several unrelated antifungal drugs[2]. Multidrug resistance is caused by the increased expression of genes that encode nonspecific drug-efflux pumps belonging to the ABC family of transporter proteins (such as Pdr5p)[3,4]. But how fungi regulate drug efflux at the level of gene transcription has remained unclear.

In *S. cerevisiae,* mutation of a single nucleotide in the genes encoding the transcription factors Pdr1p or Pdr3p could initiate multidrug resistance; these factors control the expression of several ABC efflux pumps[5–7]. Rapid and transient activation of drug-efflux genes is also observed[8] when fungi are treated with high

concentrations of antifungal drugs. Thakur *et al.*[1] find that various antifungal drugs and other xenobiotics (alien substances) can bind directly to a hydrophobic domain (XBD) in Pdr1p and Pdr3p, and that this binding leads to drug-dependent activation of drug-efflux genes.

The authors note that the direct binding of *S. cerevisiae* Pdr1p and Pdr3p to xenobiotics and their ensuing activation is functionally similar to the activation of the mammalian PXR transcription factor through its binding to various hydrophobic molecules. The similarity doesn't end there, because by binding to unrelated xenobiotics both yeast Pdr1p/Pdr3p and mammalian PXR directly activate the expression of ABC efflux pumps[9]. But until the structure of the drug-binding pocket of Pdr1p/Prd3p is determined, the functional relationships between Pdrp-like fungal transcription factors and the mammalian PXR receptor will remain hypothetical. A further complication arises from the limited evolutionary conservation of the Pdrp subfamily of transcription factors within fungal phyla. Indeed, it seems that the transcriptional mechanisms that regulate the expression of antifungal pumps in *Candida albicans*, a widespread human pathogen, are not identical to those unravelled in *S.cerevisiae*, and that *C. albicans* might use more powerful mechanisms of regulation involving genetic modification of transcription factors.
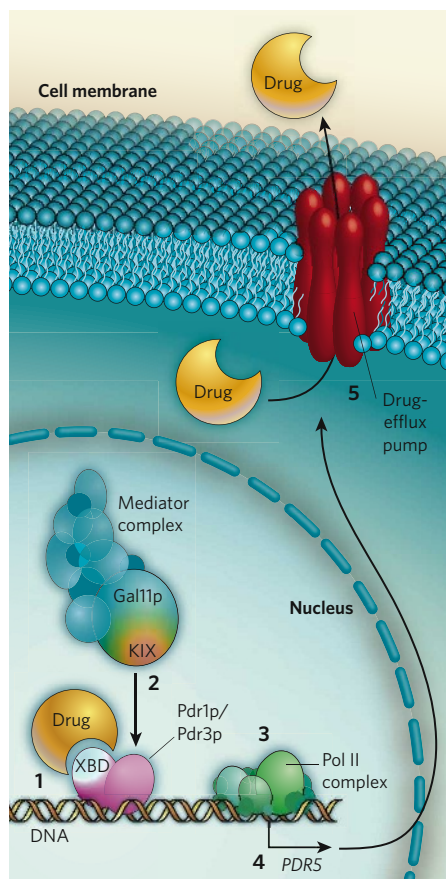
Figure 1 | **How pathogenic fungi might implement drug resistance**[1]. **1**, Within a fungal cell, an antifungal drug might bind to the XBD domain of the transcription factors Pdr1p/Pdr3p, which associate with the promoter sequence for drug-efflux genes such as *PDR5*. **2**, This would allow the binding of the Gal11p subunit of the Mediator complex through its KIX domain to the drug–Pdr1/Pdr3p–DNA complex. The Mediator complex would then facilitate (**3**) recruitment of RNA polymerase II (Pol II) and (**4**) increased transcription of *PDR5*. *PDR5* is one of a family of genes that encode drug-efflux pumps such as ABC transporters. **5**, At the cell membrane, these pumps allow the efflux of the antifungal drug, thus diminishing the fungal cell's susceptibility to the toxic compound. (Figure modified from ref. 1.)

Thakur *et al.* make another intriguing observation: deletion of the gene encoding the Gal11 protein decreases drug-dependent expression of Pdr5p and other drug-efflux pumps. Gal11p is one of the 20 subunits of the yeast Mediator complex, which recruits the RNA polymerase II machinery for the transcription of specific sets of genes. Thakur and colleagues show that the activation domain of Pdr1p/Prd3p at their carboxyl termini interacts with the KIX domain of Gal11p, and that this interaction is enhanced by antifungal drugs (Fig. 1). Moreover, they show that the yeast KIX domain is structurally similar to its mammalian counterpart, which can also bind to yeast Pdr1p. Mammalian KIX domains interact with specific transcription factors known as nuclear receptors, which affect cholesterol biosynthesis and other aspects of lipid

homeostasis. Similarly, in yeast, several aspects of lipid homeostasis depend on Pdr1p/Pdr3p (ref. 7), and this process is now shown to be regulated by the KIX domain[1].

The genomes of the non-pathogenic *S. cerevisiae* and the human pathogen *Candida glabrata* are very similar, so it is not surprising that antifungal drugs also bind to the Pdr1p protein of *C. glabrata*. Thakur and colleagues find that this interaction stimulates multidrug resistance and is mediated by one of the two Gal11p homologues in *C. glabrata*. Furthermore, when treated with antifungal drugs, nematode worms infected with *C. glabrata* in which *PDR1* and *GAL11* have been deleted lived longer. This observation not only confirms the importance of the Pdrp–Gal11p pathway in drug resistance, but also supports the use of nematode worms in studies of fungal pathogenicity.

With the discovery of the involvement of new molecular domains (XBD in Pdr1p/Pdr3p and KIX in Gal11p) in the pathway leading to antifungal-drug resistance, Thakur and colleagues argue that there is now reason to redirect the long-standing search for small molecules that inactivate resistance[10]. But finding such small molecules to prevent drug–Pdr1p/Pdr3p interaction might be difficult, because this interaction is of limited specificity. Nonetheless, the successful use of the oestrogen-receptor antagonist tamoxifen in the treatment of breast

cancer indicates that small molecules antagonizing oestrogen–PXR interaction could be found. The similarity between drug–Pdr1p/Pdr3p interaction and oestrogen–PXR interaction gives reason to believe that drug–Pdr1p/Pdr3p interaction could also be blocked with small molecules. And a broad screen, which is not restricted to the blockage of drug–Pdr1p/Pdr3p interaction but can detect any inhibitor molecule that reduces Pdr1/Pdr3p activity, has already been reported[11]. So hopes are high that we can win the battle against pathogenic fungi. ■

André Goffeau is at the Institut des Sciences de la Vie, Université catholique de Louvain, Croix du Sud 5/15, 1349 Louvain-la-Neuve, Belgium.
e-mail: andre.goffeau@scarlet.be

1. Thakur, J. K. *et al. Nature* **452,** 604–609 (2008).
2. Rank, G. H. & Bech-Hansen, N. T. *Mol. Gen. Genet.* **126,** 93–102 (1973).
3. Balzi, E., Chen, W., Ulaszewski, S., Capieaux, E. & Goffeau, A. *J. Biol. Chem.* **262,** 16871–16879 (1987).
4. Meyers, S. *et al. Curr. Genet.* **21,** 431–436 (1992).
5. Carvajal, E., van den Hazel, H. B., Cybularz-Kolaczkowska, A., Balzi, E. & Goffeau, A. *Mol. Gen. Genet.* **256,** 406–415 (1997).
6. Nourani, A., Papajova, D., Delahodde, A., Jacq, C. & Subik, J. *Mol. Gen. Genet.* **256,** 397–405 (1997).
7. DeRisi, J. *et al. FEBS Lett.* **470,** 156–160 (2000).
8. Lucau-Danila, A. *et al. Mol. Cell. Biol.* **25,**1860–1868 (2005).
9. Geick, A., Eichelbaum, M. & Burk, O. *J. Biol. Chem.* **276,** 14581–14587 (2001).
10. Lamping, E. *et al. Eukaryot. Cell* **6,** 1150–1165 (2007).
11. Kozovská, Z. & Subik, J. *Int. J. Antimicrob. Agents* **22,** 284–290 (2003).

CONDENSED-MATTER PHYSICS

Opposite of a superconductor

Rosario Fazio

Magnetism and superconductivity are caused by spontaneous ordering on a macroscopic scale. Studies of a two-dimensional superconductor reveal another striking example of such behaviour — superinsulation.

What do fireflies and iron magnets have in common? The answer is collective behaviour: fireflies synchronize their flashes, whereas the magnetism in iron results from the alignment of atomic magnetic moments. As these examples show, collective behaviour in nature can take many striking forms, but on page 613 of this issue Vinokur *et al.*[1] describe a startling example in thin films of titanium nitride. This material is usually a superconductor, but below a certain critical temperature, and in the presence of a magnetic field, it becomes a superinsulator — it reaches a state of zero electrical conductance. The conductive state of the material depends on the strength of an applied magnetic field and the thickness of the sample.

It has long been known that superconductor films can turn into insulators. But in all the previously studied systems, this has been a quantum phase transition — a change associated with quantum fluctuations close to the

absolute zero of temperature. In uniform films, in which the material is microscopically homogeneous, the transition is triggered by a drop in normal-state resistance as extra layers of atoms are added to the film[2]. The situation is different in granular films, which consist of domains of superconductivity connected by thin, insulating barriers known as Josephson junctions. The junctions behave as tiny capacitors across which superconducting charge carriers — Cooper pairs of electrons — can tunnel. The tunnelling confers charge on the junction; if the domains are small enough, then at close to absolute zero the build-up of charge is enough to prevent further current from passing. This effect has been seen both in granular films[3], and in fabricated arrays of Josephson junctions[4].

But remarkably, the state of infinite resistance observed by Vinokur *et al.*[1] is not confined to absolute zero — it extends over a range of temperatures from absolute zero up to about
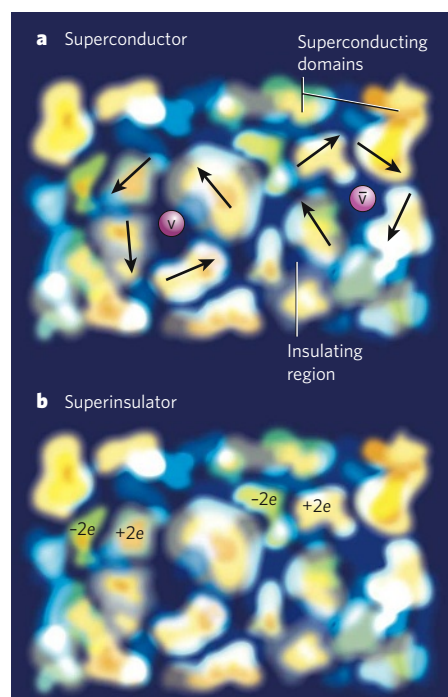
Figure 1 | Two-dimensional superconductors and superinsulators. ºa, Vinokur *et al.*[1] report a two-dimensional superconductor that can enter a superinsulating state of infinite resistance. Superconducting domains in the material are separated by insulating regions, across which charge carriers can tunnel. In the superconducting state, excited vortices of quantized magnetic flux form bound pairs (red circles, where v and v̄ indicate an opposite sense of vorticity) between the superconducting domains. Charge carriers move freely around the vortex pairs; the arrows indicate the phase of the superconductivity wavefunction around each vortex. **b,** In the superinsulating state, charge carriers and vortices have opposite roles to those in the superconducting state. Bound pairs of opposing charge form (*e* is the charge of one electron), blocking current flow, and the vortices (not shown) circulate around them.

coil acquires an additional phase proportional to the magnetic flux in the coil. The Aharonov–Casher effect is exactly the reverse — when a particle possessing a magnetic dipole moves around a charged particle, the wavefunction associated with the magnetic dipole acquires a phase proportional to that charge. So how can this sort of duality explain the bizarre conducting properties of Vinokur and colleagues' titanium-nitride film[1]?

In two-dimensional systems, the transition of a material from the normal conducting state to the superconducting state is governed by the Berezinskii–Kosterlitz–Thouless (BKT) mechanism. In the normal conducting state, single thermally excited vortices are free to move and are responsible for finite electrical resistance. But at low temperatures, they form pairs of opposite vorticity between the superconducting domains, which allows the charge carriers to move without dissipation (Fig. 1a), so that the system enters the superconducting state. For a finite voltage to exist under these conditions, a current must be applied that is sufficiently large to break the vortex pairs; the relationship between voltage and current is therefore nonlinear.

But in the superinsulating phase of Vinokur and colleagues' system[1], charges and vortices swap roles, as do current and voltage. This means that when a finite current moves around a vortex, there is a drop in voltage, in the same way that a voltage applied to charges leads to a finite current in the BKT mechanism. In the superinsulating phase, the excited charge carriers form bound pairs[9] at finite temperatures, which block electric transport (Fig. 1b). The bound charge carriers are liberated into free Cooper pairs when a sufficiently large voltage is applied. This process is itself a fascinating phenomenon, as it can be related to

very different situations, such as fluid motion in porous media or the sliding of charge density waves[10]. The fundamental relationship between superconduction and superinsulation is best illustrated in phase diagrams that plot current (for superconductors) or voltage (for superinsulators) against magnetic field and temperature (see Fig. 4 on page 615); the diagrams are essentially identical.

Vinokur and colleagues' observation[1] and theory of superinsulation are crucial advances in our understanding of the collective properties of low-dimensional systems. They also expose several issues that demand further investigation. The current understanding of this new phase relies on the model of superconductors as a network of weakly connected, superconducting puddles. But how do these puddles form, and what governs their size and dynamic behaviour? And what happens if the critical temperature for the onset of superinsulation is close to absolute zero, where quantum fluctuations play a dominant role? It is impossible to say at present, but as Vinokur *et al.* show, we should be prepared for the unexpected. ■

Rosario Fazio is at the International School for Advanced Studies (SISSA), via Beirut 2–4, I-34014 Trieste, Italy, and NEST-CNR-INFM, Pisa. e-mail: fazio@sns.it

1. Vinokur, V. M. *et al. Nature* **452,** 613–615 (2008).
2. Haviland, D. B., Liu, Y. & Goldman, A. M. *Phys. Rev. Lett.* **62,** 2180–2183 (1989).
3. Orr, B. G., Jaeger, H. M., Goldman, A. M. & Kuper, C. G. *Phys. Rev. Lett.* **56,** 378–381 (1986).
4. Geerligs, L. J. *et al. Phys. Rev. Lett.* **63,** 326–329 (1989).
5. Fisher, M. P. A. & Lee, D. H. *Phys. Rev. B* **39,** 2756–2759 (1989).
6. Fazio, R. & Schön, G. *Phys. Rev. B* **43,** 5307–5320 (1991).
7. Aharonov, Y. & Bohm, D. *Phys. Rev.* **115,** 485–491 (1959).
8. Aharonov, Y. & Casher, A. *Phys. Rev. Lett.* **53,** 319–321 (1984).
9. van der Zant, H. S. J., Geerligs, L. J. & Mooij, J. E. *Europhys. Lett.* **19,** 541–546 (1992).
10. Middleton, A. A. & Wingreen, N. S. *Phys. Rev. Lett.* **71,** 3198–3201 (1993).

70 millikelvin (for an applied magnetic field of 0.9 tesla). At higher temperatures, the titanium-nitride film behaves as a normal metal, with linear current–voltage characteristics. Below 70 millikelvin, the superinsulating phase kicks in, and no current flows unless the applied voltage exceeds a critical value. This implies that, in the superinsulating state, the Cooper pairs are somehow immobilized. When the threshold voltage is exceeded, the current dramatically increases, indicating a sudden avalanche of charge movement.

Vinokur *et al.* offer an appealing theoretical explanation for this behaviour: in superinsulators, the roles played by charges and vortices (quantum carriers of magnetic flux) in two-dimensional superconductors[5,6] are swapped. Such role reversal is not unprecedented in quantum physics, as something similar is seen in two phenomena known as the Aharonov–Bohm effect[7] and the Aharonov–Casher effect[8]. In the Aharonov–Bohm effect, the wavefunction of a charged particle crossing an area containing a

RNA INTERFERENCE

Generic block on angiogenesis

Raghu Kalluri and Keizo Kanasaki

A virtue of using small interfering RNAs as therapeutics is their exquisite specificity. But when it comes to inhibiting blood-vessel growth, it seems that they can act generically without even entering a cell.

Angiogenesis, the formation of new blood vessels from a pre-existing capillary network, is not always healthy. It accompanies the growth of cancers[1], and occurs as the 'choroidal neovascularization' that is associated with the eye disease called age-related macular degeneration[2]. To tackle this condition, clinical trials have been approved that exploit the phenomenon of RNA interference (RNAi) to shut down a signalling pathway that promotes angiogenesis. The assumption has been that the interfering RNA enters an affected cell in

some mysterious way, and has a highly specific impact in shutting down angiogenic activity.

But as Kleinman *et al.* show on page 591 of this issue[3], that assumption is not necessarily correct. Rather, these authors find that the anti-angiogenesis effect can be induced in a nonspecific manner by RNAs that vary significantly in sequence; the RNAs act through the Toll-like receptor (TLR) 3, an immune-system regulator present on the surface of many cells. These are highly provocative findings that prompt a host of questions for both

biologists and clinicians.

Gene silencing through RNAi is a promising way of treating diseases that is being widely tested in cell culture and in animals. Broadly put, it works by preventing the expression of active genes. And a particular attraction is that sequence-specific RNA agents — small interfering RNAs (siRNAs) — are highly specific for messenger RNAs of the genes they are designed to target. In the early phases, trials in patients with choroidal neovascularization revealed the efficacy of siRNAs that target the angiogenesis-promoting vascular endothelial growth factor (VEGF) or its receptor (VEGFR1). For any RNAi-based therapies to have a specific effect, it was thought that cells must take up externally administered, double-stranded siRNA. Mammalian cells do not readily take up 'naked' RNA. Nevertheless, in preclinical studies with mice, siRNAs targeting the VEGF system were injected directly into an affected eye and reduced angiogenesis.

The question remained as to how the siRNA finds its way into the cells to neutralize the gene transcripts for VEGF and VEGFR1. Kleinman and colleagues[3] set out to examine this mechanism, and to see whether, in some circumstances, siRNA might work without being taken up by the cells.

First, to test whether the effect of siRNA on angiogenesis is specific to the target sequence, the authors experimented with different nonspecific siRNAs. When injected directly into the vitreous humour — the gel between the lens and the retina — in mouse models of choroidal neovascularization, all of these nonspecific siRNAs suppressed angiogenesis. Their anti-angiogenic action was comparable to that of VEGF- and VEGFR1-specific siRNAs.

Next, Kleinman et al. explored the mechanism behind siRNA-mediated angiogenesis inhibition in the mouse models. After showing that this effect is not attributable to 'off-target' silencing by sequence-specific mismatch or generic toxicity, the authors conducted several well-controlled experiments to demonstrate that angiogenesis inhibition by any of these siRNAs does not require entry into cells. In their search for possible cell-surface mediators of inhibition, TLR3 emerged as a possible candidate. TLRs are a family of cell-surface proteins that recognize pathogen-associated molecular patterns as well as host-protein structures that convey danger signals[4,5]. TLR3 itself is a protein complex that interacts with virus-derived, double-stranded RNA and dimerizes to activate intracellular signalling pathways[6].

Kleinman et al.[3] found that all nonspecific and VEGF-system-specific siRNAs bound to TLR3 and induced angiogenesis inhibition
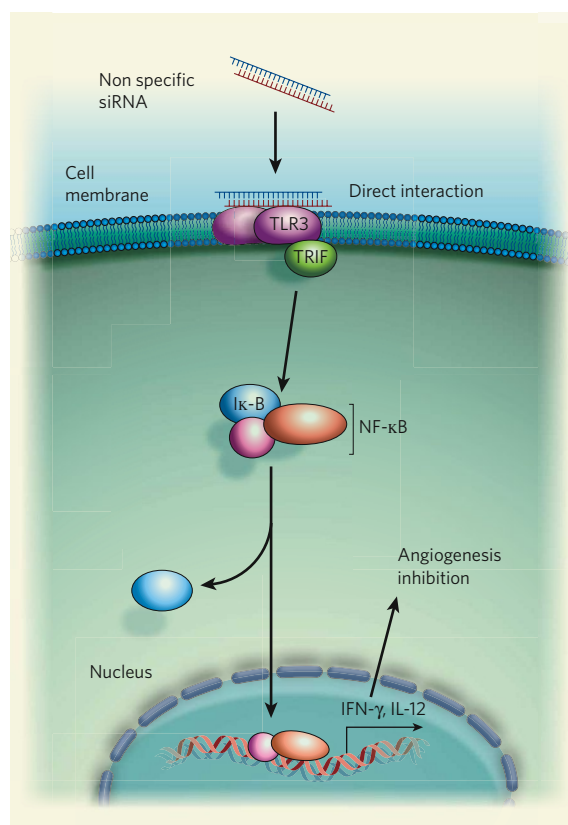
Figure 1 | Possible route of nonspecific siRNA action through Toll-like receptor 3. Kleinman et al.[3] show that any double-stranded RNA of 21 nucleotides or longer can bind directly to Toll-like receptor 3 (TLR3). This results in receptor dimerization and activation of an intracellular pathway that initially involves an adaptor protein, TRIF. Subsequently, the suppressor protein I-κB, in complex with subunits of the transcription factor NF-κB, is degraded to allow NF-κB activation of genes encoding interferon-γ (IFN-γ) and interleukin-12 (IL-12) — which, it seems, are the agents of anti-angiogenesis. In this scheme of events, therefore, inhibition of angiogenesis does not occur only through the specific action of siRNAs in suppressing messenger RNA production of vascular endothelial growth factor or its receptor.

in the mouse model of choroidal neovascularization. This effect was not observed in TLR3-deficient mice. Soluble TLR3 protein reversed the inhibitory effects of siRNAs, providing unequivocal evidence that targeted and non-targeted siRNA can interact directly with TLR3 and induce intracellular signalling that facilitates inhibition of angiogenesis.

To probe the mechanism behind TLR3-mediated inhibition of angiogenesis, Kleinman et al. studied mice lacking various components of the TLR3 signalling cascade. The results are outlined in Figure 1. After TLR3 activation and an intracellular signalling pathway, interleukin-12 and interferon-γ evidently come into play to inhibit angiogenesis. The idea that nonspecific inhibition of angiogenesis by siRNA to VEGF or VEGFR1 (the current clinical candidates) occurs through TLR3 is best supported by other results. The authors found that inhibition of angiogenesis by these siRNAs was comparable in both wild-type mice and those deficient for the gene encoding VEGF1 tyrosine kinase, but not in mice lacking the Tlr3 gene.

To have anti-angiogenic activity through TLR3, Kleinman et al. found that an siRNA had to be at least 21 nucleotides in length; truncated versions of the same siRNA did not demonstrate such activity. The authors' structural analysis shows that 21 nucleotides are required for proper docking on TLR3 and its activation through receptor dimerization on the cell surface. But although the truncated siRNAs did not activate the TLR3 receptor, when used in excess they did compete with their longer brethren for the binding sites. Further studies are required to understand this biological dichotomy.

Viral gene transcripts frequently occur in the eye, and might help to enforce the natural inhibition of angiogenesis in that setting — in other words, as they age, some people might be protected from choroidal neovascularization when exposed to viral RNA. Conversely, Kleinman et al.[3] provide preliminary data that individuals with a particular single nucleotide polymorphism in the gene encoding TLR3, and a consequent decrease in receptor function, are less susceptible to anti-angiogenic effects mediated by non-targeted and targeted siRNAs. Such genotypic variance could be exploited to identify individuals for appropriate siRNA therapies.

Finally, given their generic action as revealed by Kleinman et al.[3], the implication is that siRNAs of 21 nucleotides or more can have side effects — and thus that there should be tests for such effects before commencing clinical trials for different diseases in different organs. If an siRNA enters the cell, there will be a specific RNAi action, but there is also the possibility that nonspecific, TLR-mediated effects will occur. As well as being active in viral identification, TLR3 signalling is involved in pregnancy[7], 'immune privilege'[8] and neuronal growth[9]. Development of safe siRNA-based therapies might be more challenging than has been hoped. ■

Raghu Kalluri and Keizo Kanasaki are in the Division of Matrix Biology, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA.
e-mail: rkalluri@bidmc.harvard.edu

1. Nyberg, P., Xie, L. & Kalluri, R. Cancer Res. 65, 3967–3979 (2005).
2. Folkman, J. Nature Rev. Drug Discov. 6, 273–286 (2007).
3. Kleinman, M. E. et al. Nature 452, 591–597 (2008).
4. Akira, S. Adv. Immunol. 78, 1–56 (2001).
5. Medzhitov, R. & Janeway, C. Jr Immunol. Rev. 173, 89–97 (2000).
6. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Nature 413, 732–738 (2001).
7. Lin, Y., Liang, Z., Chen, Y. & Zeng, Y. J. Immunol. 176, 4147–4154 (2006).
8. Lang, K. S. et al. J. Clin. Invest. 116, 2456–2463 (2006).
9. Cameron, J. S. et al. J. Neurosci. 27, 13033–13041 (2007).

OBITUARY

# Arthur C. Clarke (1917–2008)

Visionary science-fiction author.

Lucky prophets become vindicated sages in their own lifetimes. Arthur C. Clarke, who died on 19 March, was the luckiest of men, becoming the most famous of science-fiction writers by writing of wonder when many were writing of despair.

Clarke was not unique in his optimism. He stood in a tradition of English futurists who have used fiction, non-fiction or both to paint their visions: H. G. Wells, J. D. Bernal, Olaf Stapledon, Freeman Dyson. They held that only science makes reliable prediction possible and the prospects for human society intelligible. Clarke believed that general laws for scientific extrapolation exist in a way that they do not in politics or economics. One such law was that "when a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong."

Clarke, the eldest of four children, was born on 16 December 1917 to farming parents in Minehead in the south-west English county of Somerset. Following service as a radar instructor and technician with the Royal Air Force in the Second World War, he honed his scientific acumen working as an editor for the academic journal *Physics Abstracts*, while earning a first-class degree in mathematics and physics at King's College London.

Humanity's leap into space confirmed Clarke's role as a seer of the possible. In the immediate postwar years, he was twice chairman of the British Interplanetary Society, one of the earliest organizations dedicated to promoting space exploration. Clarke stayed true to the society's motto, "From imagination to reality", when, in 1945, he proposed the use of satellites in geostationary orbits as communications relays. Clarke never patented the idea, but promoted it ceaselessly. Geostationary satellites have since revolutionized communications and weather forecasting, although Clarke's own dry assessment was that "my early disclosure may have advanced the cause of space communications by about fifteen minutes".

In *Profiles of the Future* (1962), his elegantly phrased "Inquiry into the Limits of the Possible", Clarke balanced his scientific knowledge with his creativity to explore more comprehensively what might be achieved within the bounds of scientific laws. Books on futurology date notoriously fast, yet this one has not, perhaps because Clarke was
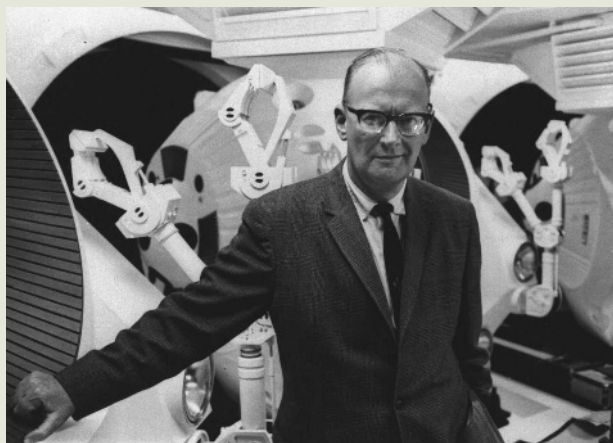
unafraid of being adventurous. And many of his predictions have come to pass: he foresaw that mobile telecommunications would mean that no one could fully escape society, even at sea or on a mountain top. He went on to describe huge electronic libraries, the breakdown of censorship and high-definition electronic screens.

Clarke favoured wise hopefulness, but saw dangers as well. His 1946 essay "The rocket and the future of warfare" anticipated the essentials of nuclear war waged with intercontinental ballistic missiles, and called for measures to avoid it. At the time, many were concerned at the prospect of nuclear conflict (the United States proposed international regulation of nuclear weapons, although the Soviet Union refused to be drawn), but few had foreseen the fateful mating of the bomb and the rocket.

Whether writing fiction or non-fiction, Clarke achieved a unique rendering of the aesthetic that combines the signal qualities of scientific endeavour: intelligence, tenacity and curiosity. His fiction has few villains, neglecting conflict and the broad spectrum of emotion. For Clarke's purposes these were pointless, even regrettable, diversions. A cool, analytical tone, springing from a pure, dispassionate statement of facts and relationships, pervades his writing. But the result is never cold, and indeed is often aphoristically witty. On religion: "I don't believe in God, but I'm very interested in her." Of space and politics: "There is hopeful symbolism in the fact that flags do not wave in a vacuum." On progress: "New ideas pass through three periods: it can't be done; it probably can be done, but it's not worth doing; I knew it was a good idea all along."

Undoubtedly Clarke's greatest success was his co-authorship, with Stanley Kubrick, of

the film and novel *2001: A Space Odyssey*. The film's technological futurism, grounded in hard science, together with its iconic, mystical opening, made it a cultural benchmark and the embodiment of Clarke's assertion that "whatever other perils humanity may face in the future that lies ahead, boredom is not among them". Kubrick observed of Clarke: "He has the kind of mind of which the world can never have enough, an array of imagination, intelligence, knowledge and a quirkish curiosity, which often uncovers more than the first three qualities." Clarke later relished the story of an immigration official threatening to bar his entry to the United States until he had explained the ending of *2001*.

Clarke moved permanently to Sri Lanka in 1956 to pursue one of his lifelong passions, scuba diving. He continued to dive even after he was partially confined to a wheelchair in the 1980s: he felt, he said, "perfectly operational underwater". He founded Asia's first diving shop there, and restored the business after the 2004 tsunami. Sri Lanka gave Clarke creative isolation, but he was an early citizen of the global village, travelling and later keeping in touch with friends, colleagues and fans through daily e-mails. Surrounded by tropical cuisine, he never gave up his taste for a steady diet of English food, roast beef and Yorkshire pudding being a perennial favourite.

His home in Sri Lanka contained a room he called the Ego Chamber. Amid the many awards stored there is a tribute from Apollo 11 astronaut Neil Armstrong that epitomizes Clarke's peculiar contribution to science: "To Arthur — who visualized the nuances of lunar flying before I experienced them."

Up to the last, Clarke viewed humanity's ability to further expand its realm of experience, and so survive and prosper, with signature wry confidence. As he told a convention of the European Mars Society in 2002 via video-link from Sri Lanka: "Whether we become a multi-planet species with unlimited horizons, or are for ever confined to Earth, will be decided in the twenty-first century amid the vast plains, rugged canyons and lofty mountains of Mars."

Gregory Benford

Gregory Benford is in the Department of Physics and Astronomy, University of California, Irvine, Irvine, California 92697-4575, USA. He is also an author whose best-known novel is *Timescape*.
e-mail: xbenford@aol.com

MOLECULAR CANCER DIAGNOSTICS

nature insight
MOLECULAR CANCER DIAGNOSTICS

Cover illustration
Fluorescence-mediated tomography image showing angiogenesis (red) and protease activity (dark pink) in lung cancer. (Courtesy of R. Weissleder)

A myriad of differences in the biological make-up of cancers compared with their healthy-tissue counterparts have been catalogued. Large-scale analyses have detailed the heterogeneity of these changes at the genetic and epigenetic levels, as well as in gene-expression profiles and signal-transduction networks in a wide range of cancers. This knowledge has led to the discovery and regulatory approval of drugs that target cancer-specific pathways.

Even as the first genetic alterations in human cancers were unravelled, scientists proposed that differences between individuals and between the various types of cancer might necessitate personalized therapies. But the drug-discovery process has proven difficult, and the rate of the approval of targeted cancer therapies remains woefully low. A key challenge and priority in cancer research is therefore to identify molecular biomarkers that could be used to improve early diagnosis, as well as to guide prognosis and the design and monitoring of new therapeutic avenues.

This Insight highlights key approaches for the discovery and validation of such biomarkers (or diagnostics) at the level of DNA, RNA and proteins. It also looks at how non-invasive imaging technologies and serum biomarkers can be used to monitor early responses to treatment and guide the therapeutic course.

Translating these advances into personalized cancer care will entail challenges far beyond the scientific discovery and validation strategies. We hope that the articles in this Insight not only bring together key aspects of the translational research into cancer biomarkers but also draw attention to associated issues such as trial design, tissue collection and regulatory requirements.

Barbara Marte, Alex Eccleston, Deepa Nath, Senior Editors

nature insight

The cancer biomarker problem

Charles L. Sawyers[1]

Genomic technologies offer the promise of a comprehensive understanding of cancer. These technologies are being used to characterize tumours at the molecular level, and several clinical successes have shown that such information can guide the design of drugs targeted to a relevant molecule. One of the main barriers to further progress is identifying the biological indicators, or biomarkers, of cancer that predict who will benefit from a particular targeted therapy.

In the past decade, there have been considerable improvements in the way that human tumours are characterized. Knowledge of cancer at the molecular level has therefore increased greatly, and this has catalysed a shift towards using targeted therapies for cancer. However, there has been much less progress in the development of clinical tools to determine which patients are most likely to benefit from particular targeted therapies. The articles that follow in this Insight summarize the tremendous advances that have been made in the molecular characterization of tumours, and this overview outlines the considerable challenges that remain before these advances can have the maximum clinical impact.

Much progress has recently been made in defining tumour cells in terms of two important features: DNA copy number (that is, the number of copies of each genetic region or chromosome, a property that is commonly aberrant in cancer cells), as discussed by Lynda Chin and Joe Gray (see page 553); and patterns of gene expression, as reviewed by

Laura van 't Veer and René Bernards (see page 564). Advances in DNA-microarray technology have made it possible to define almost completely the chromosomal gains and losses in individual tumours and the resultant changes in gene expression, at very high resolution in a robust and reproducible manner (Box 1). When coupled with focused resequencing of specific genes to look for point mutations that are not detectable by DNA-microarray analysis, it is now possible to characterize individual human cancers in unprecedented molecular detail.

The impact of this improvement in characterization, initially on the conduct of clinical trials and subsequently on clinical practice, is potentially enormous. Consider, for example, the sudden interest of doctors and the general public in single-nucleotide-polymorphism-based genotyping to find alleles associated with an increased risk of developing certain diseases; this occurred on the heels of a remarkable series of discoveries of 'risk alleles' for cancer, diabetes and other diseases over the past year[1]. The uptake of molecular tools in oncology practice is similarly promising. It is more difficult logistically, however, because tumour tissue, which can be difficult to gain access to, is required rather than germline DNA, which can be obtained from any cell. Therefore, it is important to be able to assess tumours non-invasively. Several technologies offer the promise of analysing tumours comprehensively at the molecular level, both quantitatively and qualitatively, without subjecting patients to multiple clinical interventions to obtain tissue (Box 2). In particular, the use of proteomic technologies to analyse cancer-associated changes in serum proteins is discussed by Samir Hanash, Sharon Pitteri and Vitor Faca (see page 571), and the molecular imaging of tumours *in situ* is reviewed by Ralph Weissleder and Mikael Pittet (see page 580).

Collectively, the fields of genomics, proteomics and molecular imaging have matured to a level at which they are ripe for clinical exploitation. But there are considerable barriers to broad implementation of these technologies in the clinic. The challenge is discovering cancer biomarkers. Although there have been clinical successes in targeting molecularly defined subsets of several tumour types — such as chronic myeloid leukaemia, gastrointestinal stromal tumour, lung cancer and glioblastoma multiforme —using molecularly targeted agents, the ability to apply such successes in a broader context is severely limited by the lack of an efficient strategy to evaluate targeted agents in patients. The problem mainly lies in the inability to select patients with molecularly defined cancers for clinical trials to evaluate these exciting new drugs. The solution requires biomarkers that reliably identify those patients who are most likely to benefit from a particular agent. In this overview, I consider the complex set of barriers — logistical, scientific and commercial — that impede progress, and I argue for a public–private consortium approach to cancer biomarker discovery.

> **Box 1 | Technologies for characterizing tumours**
>
> Molecular alterations in tumours can be uncovered by using technologies that assess changes in the content or sequence of DNA, its transcription into messenger RNA or microRNA, the production of proteins or the synthesis of various metabolic products. Below is a partial list of the various types of information that can be obtained about tumours and some of the technologies that are used to make those assessments.
>
> **DNA copy-number assessment**
> • Comparative genome hybridization to DNA microarrays
>
> **Mutation screening**
> • DNA sequencing
> • Mass-spectrometry-based genotyping
> • Mutation-specific PCR
>
> **Gene-expression profiling**
> • DNA microarrays
> • Multiplex PCR
>
> **MicroRNA-expression profiling**
> • DNA microarrays
> • Multiplex PCR
>
> **Proteomic profiling**
> • Mass spectrometry
>
> **Phosphoproteomic profiling**
> • Mass spectrometry after immunoprecipitation with phosphotyrosine-specific antibodies
>
> **Metabolomic profiling**
> • Mass spectrometry

[1]Howard Hughes Medical Institute, Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA.

Types of cancer biomarker

The topic of cancer biomarkers is broad, encompassing multiple disciplines, and it has been the focus of several recent reviews, workshops and planning committees[2]. Although most people agree on the scope of the problem, there is no consensus on a strategy to move forward and even scepticism about the ultimate usefulness of cancer biomarkers[3,4]. Here, I discuss three types of cancer biomarker — prognostic, predictive and pharmacodynamic — each of which can aid in the rational development of anticancer drugs (Fig. 1). Prognostic biomarkers allow the natural course of an individual cancer to be predicted, distinguishing 'good outcome' tumours from 'poor outcome' tumours, and they guide the decision of whom to treat (or how aggressively to treat). Notable recent examples include breast-cancer gene-expression signatures — marketed for clinical use as Oncotype DX (Genomic Health), MammaPrint (Agendia) and the H/I test (AviaraDx) — that estimate the probability of the original breast cancer recurring after it has been resected (that is, surgically removed). These multigene-expression tests can now be used to decide who should receive systemic therapy to eliminate any remaining tumour cells (that is, adjuvant therapy) after surgery, to reduce the risk of relapse.

Predictive (or response) biomarkers differ in that they are used to assess the probability that a patient will benefit from a particular treatment. Patients with breast cancer in which the gene *ERBB2* (also known as *HER2* or *NEU*) is amplified (that is, extra copies are present) benefit from treatment with trastuzumab (Herceptin), whereas when the gene encoding the oestrogen receptor is expressed by the tumour, the patients respond to treatment with tamoxifen instead. Similarly, patients who have leukaemia with the *PML–RARA* translocation respond to all-*trans* retinoic acid, and those with the Philadelphia chromosome (which contains the *BCR–ABL* fusion gene) respond to imatinib mesylate (Gleevec or Glivec). Biomarkers for leukaemia have traditionally been assessed by using routine cytogenetic analysis, but additional predictive information can be gained by using genotype-based analysis. For example, in patients with chronic myeloid leukaemia who develop resistance to imatinib mesylate, distinct mutations in the genetic region encoding the kinase domain of BCR–ABL predict differential sensitivity to the newer ABL inhibitors dasatinib and nilotinib[5]. In addition, mutations in the genetic region encoding the kinase domain of the epidermal growth-factor receptor (EGFR) predict the sensitivity of lung tumours to erlotinib or gefitinib[6]. Conversely, distinct mutations in *KRAS* predict that patients with lung cancer will fail to respond to these inhibitors and that patients with colon cancer will fail to respond to therapy with EGFR-specific antibody[7,8]. And, in glioblastoma multiforme, distinct mutations in the genetic region encoding the extracellular domain of EGFR predict sensitivity to EGFR inhibitors but only in cases in which the tumour-suppressor protein PTEN is also intact[9].

Pharmacodynamic biomarkers measure the near-term treatment effects of a drug on the tumour (or on the host) and can, in theory, be used to guide dose selection in the early stages of clinical development of a new anticancer drug. In cytotoxic chemotherapy, the dose that is used to determine antitumour activity in phase II clinical trials is usually the maximum tolerated dose, discovered in a phase I dose-escalation study. But this might be a less relevant end point for drugs that have been optimized to bind to a specific molecular target. An alternative way to determine an appropriate dose is to measure the impact of the drug on its target across a range of doses (known as a target engagement study) and then to select a dose for phase II clinical trials on the basis of the magnitude of target modulation. For example, imatinib mesylate has been shown to block the protein-kinase activity of BCR–ABL in the tumour cells of patients with chronic myeloid leukaemia at the same doses that induce clinical remission, which are well below those associated with toxicity. The utility of pharmacodynamic biomarkers might also extend beyond the clinical trial phase of drug development. Recently, the magnitude of BCR–ABL kinase activity inhibition was found to correlate with clinical outcome, possibly justifying the personalized selection of drug dose based on the results of target engagement assays[10].

> **Box 2 | Non-invasive strategies for the molecular profiling of cancer**
>
> Several technologies offer the promise of detecting cancer without the need for carrying out a biopsy or a surgical procedure and (when cancer is present) of allowing certain molecular studies of tumour cells. The examples listed are discussed here, in the accompanying Insight articles or in the cited references.
>
> - Analysing circulating tumour cells[15,17] (see page 580)
> - Carrying out mutation-specific PCR on circulating DNA[26]
> - Using proteomic approaches to study serum or plasma (see page 571)
> - Imaging tumours *in situ* at the molecular level (see page 580)
> - Assessing autoantibodies specific for tumour cells[27]

Studying biomarkers in patients with solid tumours

Predictive and pharmacodynamic biomarkers are increasingly being used in clinical trials of leukaemia drugs. But there has been much less uptake in clinical trials for treating solid tumours, because it is challenging to gain access to tumour tissue so that biomarkers can be measured. Unlike leukaemia, in which large numbers of tumour cells are present in the peripheral blood, the only time at which access to solid tumour tissue is guaranteed is at diagnosis, when the tumour is biopsied or resected. Although this approach might be sufficient to study prognostic biomarkers, it severely limits the application of predictive and pharmacodynamic biomarkers, because these measurements are ideally carried out concurrently with treatment. In addition, experimental drugs are typically evaluated in patients with late-stage disease, who do not routinely undergo additional tumour biopsies.

Despite much discussion on this topic, most clinical trials in patients with solid tumours do not include provisions to obtain additional tissue samples. The reasons for this include concerns about slower patient enrolment in trials and insufficient research infrastructure to obtain and process tumour samples for molecular analysis. The reluctance of the clinical trial community to force the issue by designing studies that require additional tissues may be short-sighted. Successful 're-biopsy' studies have been carried out and have resulted, for example, in the discovery that patients with lung cancer who relapse during treatment with EGFR inhibitors have acquired mutations in the genetic region encoding the kinase domain of EGFR since the initial biopsy[11].

An alternative to the re-biopsy approach is to evaluate the experimental drug in patients who are already scheduled to undergo a surgical procedure as part of their treatment regimen. In this case, the treatment is administered for a brief period (days to weeks) before surgery, and the effect of the intervention on the tumour is assessed by comparing the surgically removed sample with the diagnostic biopsy sample (obtained before treatment). In one example of this approach, the proportion of proliferating tumour cells (as determined by staining cells with an antibody specific for the antigen Ki-67) was measured in patients with breast cancer who were undergoing resection after 2 weeks of neo-adjuvant
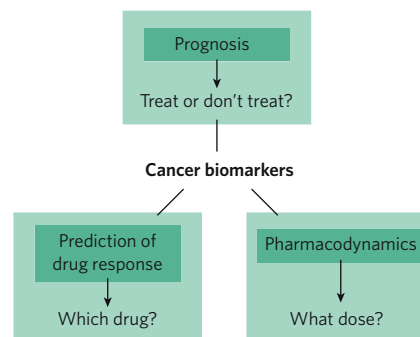

Figure 1 | Types of biomarker. Cancer biomarkers can be used for prognosis: to predict the natural course of a tumour, indicating whether the outcome for the patient is likely to be good or poor (prognosis). They can also help doctors to decide which patients are likely to respond to a given drug (prediction) and at what dose it might be most effective (pharmacodynamics).
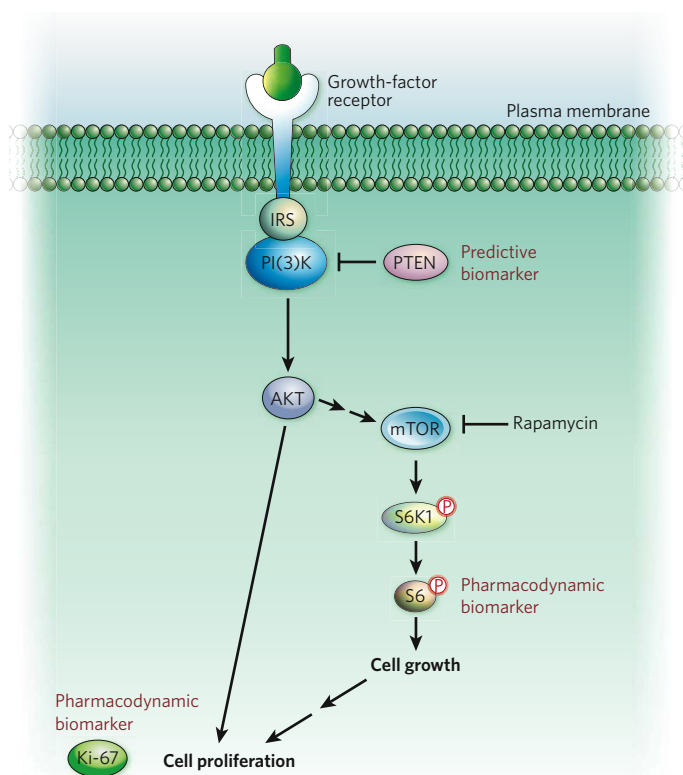
Figure 2 | Biomarkers in the PI(3)K–PTEN–mTOR pathway. The phosphatidylinositol-3-OH kinase (PI(3)K)–PTEN–mTOR signalling pathway is aberrantly activated in many tumours, leading to dysregulation of cell growth and proliferation. Activation of the pathway can be assessed by biomarkers such as loss of PTEN mRNA or protein production in tumour tissue. Biochemical inhibition of mTOR by rapamycin can be assessed by biomarkers such as the abundance of the phosphorylated form of the ribosomal protein S6, and its therapeutic effects on tumour cells can be assessed by the proliferation marker Ki-67. IRS, insulin-receptor substrate; S6K1, ribosomal protein S6 kinase, 70-kDa, polypeptide 1.

(that is, presurgical) hormonal therapy. In most patients, the proportion of tumour cells that were Ki-67$^+$ was smaller than in the pretreatment biopsy sample, and the magnitude of this reduction correlated with progression-free survival (that is, the length of time in which the individual's disease did not worsen)[12]. In another example, patients with glioblastoma multiforme received the mTOR inhibitor rapamycin (Fig. 2) for 1 week before a salvage resection (which is often carried out to reduce symptoms associated with tumour recurrence after initial surgery) and then post-operatively until evidence of tumour progression was obtained by magnetic resonance imaging[13]. On the basis of preclinical studies showing that loss of *PTEN* expression made tumours more sensitive to rapamycin[14], eligibility for the study was restricted to patients whose tumours showed loss of PTEN production, as determined by analysing the tissue from the initial surgery. The proportion of Ki-67$^+$ cells was substantially reduced in the tumours of half of the patients, and this reduction was correlated with the extent to which mTOR's kinase activity was inhibited, as measured by staining for the phosphorylated form of the ribosomal protein S6 (which is downstream of mTOR in the intracellular signalling pathway) (Fig. 2). Therefore, analysing tumour tissues for two biomarkers — Ki-67 and phosphorylated S6 — showed the importance of documenting target inhibition for guiding patient-specific dose selection.

Because the neo-adjuvant trial design and the re-biopsy trial design are challenging, there is great interest in developing tools that can gain access to tumour cells non-invasively, namely by a blood test. This process would help to gather the molecular information that is required to make informed decisions during clinical development of a new drug and, after regulatory approval, to identify those patients who are most likely to benefit from the drug. Much effort is focused on whether this

information can be gained by studying the very small proportion of tumour cells that circulate in the blood or the cancer-associated proteins that are secreted or shed into the blood.

Recently, it has become clear that these rare circulating tumour cells (CTCs) are present in the blood of many patients with cancer and can be recovered by immunoaffinity purification, using antibodies specific for cell-surface proteins restricted to epithelial cells. For patients with breast cancer, the number of CTCs per 7.5 ml of blood is a prognostic biomarker[15], and there is great interest in determining whether a decrease in CTC number concomitant with treatment predicts long-term benefit. In prostate cancer, CTCs have been shown to contain the same genetic alterations as the primary cancer (such as amplification of the gene encoding the androgen receptor), indicating that CTCs could provide a window onto the tumour genome[16]. Despite these promising reports, CTCs can be detected reliably only in patients with advanced metastatic disease, and the number of CTCs is extremely small — often only 5–10 cells per 7.5 ml of blood. But evidence obtained by using microfluidic technologies indicates that CTCs might be present in a much larger proportion of patients (including those at earlier stages of disease) and in about 10–100-fold greater quantity[17].

Instead of studying the tumour cells themselves, it might be possible to characterize the molecular composition of a tumour indirectly, by sampling the blood and searching for alterations in the serum proteins. The idea of using blood to track cancer growth is well established in terms of measuring changes in the abundance of proteins that are secreted by the tumour (for example, prostate-specific antigen for prostate cancer, carcinoembryonic antigen for multiple cancers, CA125 for ovarian cancer and α-fetoprotein for liver cancer and testicular cancer). But this technique is not broadly applicable because of the paucity of known biomarkers and because most of the markers are organ specific rather than tumour specific. Mass-spectrometry-based proteomic technologies offer the promise of a genome-scale search for tumour-specific serum biomarkers and could transform the early detection and molecular characterization of cancers through non-invasive means. The initial enthusiasm for cancer-specific serum proteomics was tempered by problems with reproducibility. But progress in overcoming limitations in the sensitivity of detection, in the ability to make quantitative measurements and in the standardization of sample collection has led to increased confidence in data collection, and several large-scale, collaborative serum proteomic programmes are now underway (see page 571).

Discovering predictive biomarkers

Moving beyond the technical considerations of collecting tumour tissue, there is considerable debate about precisely which measurements will be most informative for predicting how a patient will respond to treatment. It is clear that gene-expression signatures have value as prognostic biomarkers (see page 564), but their value for predicting responses to particular treatments is less convincing. One reason is that the tissue collection associated with most treatment studies is incomplete, so the number of samples available is often too small to allow a formal evaluation of the hypothesis. As mentioned earlier, prognostic biomarker studies analyse tissue obtained at diagnosis, whereas studies assessing treatments (in which predictive or pharmacodynamic biomarkers could be measured) are typically carried out in patients with advanced disease, who do not routinely undergo surgery for additional tissue samples to be collected. Therefore, the 'data-driven' (unbiased) approach to prognostic biomarker discovery that is advocated by van 't Veer and Bernards (that is, surveying the entire genome rather that working from a hypothesis about a candidate biomarker) cannot be implemented for predicting treatment responses without overcoming enormous logistical challenges. An alternative strategy might be to search for candidate predictive gene-expression signatures in preclinical models (such as cell-line and animal models) and then to validate these in the clinic, thereby reducing the number of patients required for tissue collection[18]. But early reports of success using this approach have been challenged[19].

By contrast, the genotyping of tumour DNA has been found to be more useful for predicting responses to treatment, but the path to broader

clinical application of this technique remains unclear. The lung cancer, glioblastoma multiforme and chronic myeloid leukaemia examples discussed earlier show the value of knowing the 'mutation status' of the genes encoding the targets of protein-kinase inhibitors, but these studies also reveal the complex role of secondary mutations in predicting responses to treatment. For example, amplification of the gene encoding the drug target (in the absence of mutation) might be associated with sensitivity to treatment[6]. Therefore, predictive biomarkers must incorporate DNA copy-number assessment together with mutation detection. For most protein-kinase inhibitors, resistance is associated with secondary mutations in the gene encoding the drug target, so these mutations must be considered as an explanation for treatment failure. Tests that focus exclusively on the common oncogenic mutations in the gene encoding the drug target are therefore unlikely to detect drug-resistant variants that contain these secondary mutations. Furthermore, these drug-resistant alleles might be present in only a small proportion of the tumour cells initially and would be not be detected unless highly sensitive techniques, such as single-molecule sequencing, were used[20]. Also, mutations in additional genes such as *RAS* or *PTEN* can dampen sensitivity to inhibitors of the drug target that has mutated (as was the case in the lung cancer and glioblastoma multiforme examples discussed earlier). Therefore, a predictive biomarker test must incorporate these additional variables. Finally, patients who do not carry alterations in the gene encoding the drug target can have clinical responses, as in the case of treating lung cancer with EGFR inhibitors. Such tumours might be sensitive to EGFR inhibitors as a result of mutations in genes (either known or unknown) that regulate the EGFR pathway, but this association can be discovered only by more extensive sequencing.

The value of tumour-DNA genotyping is clear, but the challenge now is to decide how broadly and deeply to genotype. All of the measurements described here can be made with available technologies, but costs escalate quickly depending on the scale: that is, on the number of genes analysed by sequencing and the level of sensitivity required to carry out a thorough analysis. The data-driven (unbiased) approach analogous to that used for gene-expression profiling would require complete (or at least exon-focused) resequencing of the entire cancer genome. Such proposals cannot be considered today because of the prohibitive cost, but this barrier might soon disappear with the advent of improved sequencing technologies. Even then, the computational infrastructure required for data analysis and comparisons across tumour populations will be formidable. In the interim, it seems logical to use a more hypothesis-driven approach, such as genotyping patients for known cancer-associated mutations by using platforms that can easily be expanded to accommodate new discoveries. For example, the presence or absence of 238 cancer-associated mutations in 1,000 tumour-derived cell lines and archived tumour tissue samples has been determined using mass-spectrometry-based genotyping[21]. The value of such assessments in a clinical setting remains to be determined.

Because most of the anticancer agents in development inhibit targets in specific molecular pathways, another option would be to focus on pathway-specific biomarkers. The activation state of many pathways can be assessed by using antibodies that recognize downstream substrates in the pathway only when they are in their 'activated' form. The most common examples are phospho-specific antibodies, which recognize the substrates of various protein kinases after the kinases have phosphorylated them. Some of these antibodies have been used to document the inhibition of protein-kinase activity in clinical trials[13]. So far, pathway activation has been studied with only a few antibodies in the context of a known genetic lesion. But more global approaches such as those based on mass spectrometry could be used as an initial screening step to find tumours that are appropriate for focused sequencing, which could then identify the causative genetic lesion. Indeed, in global surveys of the phospho-tyrosine-containing proteome in lung cancer, previously unidentified protein-tyrosine-kinase fusion proteins were found[22]. Extending this approach to other post-translational modifications, such as the acetylated or ubiquitylated proteome, is feasible as soon as robust antibodies that recognize the modifications of interest are available. Although these approaches rely on the direct assessment of pathway substrates to measure pathway activation, more indirect approaches might be possible using gene-expression signatures. The underlying concept — that pathway activation is associated with a specific gene-expression signature — has been shown for cell lines engineered to express specific oncogenes and for tumours with specific pathway-activating alterations such as loss of PTEN production[23–25].

Commercializing cancer biomarkers

Even if the logistical and scientific issues in cancer biomarker discovery can be overcome, there is concern that the commercial incentives to develop these complex assays for broad clinical use might not be in place. The process is expensive and lengthy because the biomarker must be identified, an assay that measures the biomarker reliably in clinical samples must be developed (validation) and the capacity of the biomarker to make a clinical distinction must be demonstrated (qualification). One strategy is to pair the diagnostic test with the therapeutic agent, an idea that is best illustrated by the development of a standardized immunohistochemical assay for ERBB2 protein (HercepTest; DAKO). This test identifies which patients with breast cancer are most likely to benefit from treatment with the ERBB2-specific antibody trastuzumab. In this model (sometimes referred to as the Dx/Rx model), the incentive for identification, validation and qualification of a predictive biomarker (all of which are essential to obtain drug approval) lies with the drug manufacturer, who therefore drives its commercial development, often in collaboration with a molecular diagnostics company. One huge challenge is that discovery of the biomarker and clinical testing of the drug are interdependent and move forward in parallel. Therefore, crucial decisions about biomarker-driven selection of patients for the phase III registration trial (which is required for drug approval) must often be made before the utility of the biomarker has been shown.

An alternative to the Dx/Rx model is to use pathway-based biomarkers to classify cancers into categories that are more appropriately matched to the many pathway-focused inhibitors in development. In this model, tumours would be categorized at diagnosis into distinct molecular subtypes, similarly to the current practice of karyotyping for chromosomal alterations cancer cells from all patients who are diagnosed with leukaemia. The challenge lies in discovering the biomarkers that will provide the best classification. As is the case for the Dx/Rx model, the incentive to discover and validate these biomarkers seems to lie with the companies developing pathway-specific drugs. But the scale of the research effort required to discover and commercialize pathway-based biomarkers is enormous and is probably beyond the capacity of most companies. Furthermore, the first company to succeed in defining predictive pathway biomarkers will make the process much simpler for its competitors, because the method of pathway classification used to gain approval will be in the public domain.

Another consideration is the question of biomarker ownership. The discovery of predictive biomarkers is likely to be a gradual process, building on the collection of large data sets from preclinical studies and clinical trials carried out by pharmaceutical companies in collaboration with academic partners. Furthermore, the biomarkers that ultimately prove most useful in the clinic are likely to include a suite of measurements that are modified over time as further clinical evaluation improves their predictive power. In this case, when (if ever) is it appropriate for any party to claim ownership of the biomarker? Will the filing of patents on various components of a multi-parameter biomarker impede commercial development? And is patent protection even essential for molecular diagnostic companies to enter the cancer biomarker arena? If patent protection is required, then one option to avoid the patenting of biomarkers themselves is to patent the tools that were developed to measure the biomarkers.

Regulatory authorities, although generally focused on approval of anticancer drugs, also have a crucial role in biomarker development, because validation and qualification of the predictive biomarker is required for drug approval in the Dx/Rx model. This level of regulatory endorsement provides the evidence, often demanded by health-care

payers (such as insurance companies and government agencies), to justify the reimbursement of patients and hospitals for the biomarker test, and it is a powerful commercial incentive for molecular diagnostic companies to enter the biomarker arena. But if the drug-development community moves away from the Dx/Rx model, the role of regulatory endorsement is less clear. In the United States, a biomarker test can be marketed without a formal demonstration of its clinical value if the technical aspects of the biomarker measurement are certified under Clinical Laboratory Improvement Amendments (CLIA). CLIA certification has encouraged the development of numerous highly innovative molecular assays, but few are put to the more rigorous test of prospective clinical qualification because of the time and expense required. At present, the price charged for these tests is often reimbursed by health-care payers, but growing pressure to reduce health-care costs will lead to greater scrutiny. With the move to more expensive assays that survey tumours for mutations or pathway activation, it will be crucial that the regulatory strategy used to approve these assays inspires health-care payers to be confident about reimbursement so that an overly burdensome initial proof of clinical value is not required. Similar to the current process of provisional drug approval, a strategy of provisional biomarker approval that encourages small, innovative molecular diagnostic companies to enter the marketplace can be envisaged.

A public–private biomarker consortium

The 'omic' technologies reviewed in this Insight are poised to launch a comprehensive approach to cancer biomarker discovery. The scale of this endeavour, which includes preclinical studies and large clinical trials, is considerable, making it expensive. Collaboration with the pharmaceutical industry is essential because experimental anticancer drugs are an essential reagent for biomarker discovery experiments. Many cancer biomarkers will be broadly applicable (for example, they will not be restricted to predicting the response to a single drug), so a collaborative, precompetitive partnership with industry is warranted. Similar to government-sponsored projects such as the Human Genome Project and The Cancer Genome Atlas, early results from collaborative biomarker discovery projects should be released into the public domain to encourage further study and to avoid downstream intellectual-property disputes that could delay commercialization efforts. It is time to establish a consortium approach using a public–private partnership model to solve the cancer biomarker problem. All the stakeholders — patients, doctors, pharmaceutical and molecular diagnostic companies, regulatory agencies and heath-care payers — stand to benefit. ∎

1. Hunter, D. J., Khoury, M. J. & Drazen, J. M. Letting the genome out of the bottle — will we get our wish? *N. Engl. J. Med.* **358,** 105–107 (2008).
2. Committee on Developing Biomarker-Based Tools for Cancer Screening, Diagnosis, and Treatment. *Cancer Biomarkers: the Promises and Challenges of Improving Detection and Treatment* (eds Nass, S. J. & Moses, H. L.) (National Academies Press, Washington DC, 2007).
3. Ratain, M. J. & Glassman, R. H. Biomarkers in phase I oncology trials: signal, noise, or expensive distraction? *Clin. Cancer Res.* **13,** 6545–6548 (2007).
4. Carroll, K. J. Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology. *Pharm. Stat.* **6,** 253–260 (2007).
5. Shah, N. P. *et al.* Sequential ABL kinase inhibitor therapy selects for compound drug-resistant *BCR-ABL* mutations with altered oncogenic potency. *J. Clin. Invest.* **117,** 2562–2569 (2007).
6. Sharma, S. V., Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nature Rev. Cancer* **7,** 169–181 (2007).
7. Pao, W. *et al.* *KRAS* mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. *PLoS Med.* **2,** e17, doi:10.1371/journal.pmed.0020017 (2005).
8. Khambata-Ford, S. *et al.* Expression of epiregulin and amphiregulin and *K-ras* mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *J. Clin. Oncol.* **25,** 3230–3237 (2007).
9. Mellinghoff, I. K. *et al.* Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. *N. Engl. J. Med.* **353,** 2012–2024 (2005).
10. Shah, N. P. *et al.* Potent transient inhibition of BCR–ABL by dasatinib leads to complete cytogenetic remissions in patients with chronic myeloid leukemia: implications for patient management and drug development. *Blood* **108,** abstr. 2166 (2006).
11. Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med.* **2,** e73, doi:10.1371/journal.pmed.0020073 (2005).
12. Dowsett, M. *et al.* Prognostic value of Ki67 expression after short-term presurgical endocrine therapy for primary breast cancer. *J. Natl Cancer Inst.* **99,** 167–170 (2007).
13. Cloughesy, T. F. *et al.* Antitumor activity of rapamycin in a phase I trial for patients with recurrent PTEN-deficient glioblastoma. *PLoS Med.* **5,** e8, doi:10.1371/journal.pmed.0050008 (2008).
14. Neshat, M. S. *et al.* Enhanced sensitivity of PTEN-deficient tumors to inhibition of FRAP/mTOR. *Proc. Natl Acad. Sci. USA* **98,** 10314–10319 (2001).
15. Cristofanilli, M. *et al.* Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N. Engl. J. Med.* **351,** 781–791 (2004).
16. Shaffer, D. R. *et al.* Circulating tumor cell analysis in patients with progressive castration-resistant prostate cancer. *Clin. Cancer Res.* **13,** 2023–2029 (2007).
17. Nagrath, S. *et al.* Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **450,** 1235–1239 (2007).
18. Potti, A. *et al.* Genomic signatures to guide the use of chemotherapeutics. *Nature Med.* **12,** 1294–1300 (2006).
19. Coombes, K. R., Wang, J. & Baggerly, K. A. Microarrays: retracing steps. *Nature Med.* **13,** 1276–1277 (2007).
20. Thomas, R. K. *et al.* Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Med.* **12,** 852–855 (2006).
21. Thomas, R. K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nature Genet.* **39,** 347–351 (2007).
22. Rikova, K. *et al.* Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **131,** 1190–1203 (2007).
23. Mehrian-Shai, R. *et al.* Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer. *Proc. Natl Acad. Sci. USA* **104,** 5563–5568 (2007).
24. Saal, L.H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl Acad. Sci. USA* **104,** 7564–7569 (2007).
25. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–357 (2006).
26. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl Acad. Sci. USA* **102,** 16368–16373 (2005).
27. Wang, X. *et al.* Autoantibody signatures in prostate cancer. *N. Engl. J. Med.* **353,** 1224–1235 (2005).

Translating insights from the cancer genome into clinical practice

Lynda Chin[1] & Joe W. Gray[2]

Cancer cells have diverse biological capabilities that are conferred by numerous genetic aberrations and epigenetic modifications. Today's powerful technologies are enabling these changes to the genome to be catalogued in detail. Tomorrow is likely to bring a complete atlas of the reversible and irreversible alterations that occur in individual cancers. The challenge now is to work out which molecular abnormalities contribute to cancer and which are simply 'noise' at the genomic and epigenomic levels. Distinguishing between these will aid in understanding how the aberrations in a cancer cell collaborate to drive pathophysiology. Past successes in converting information from genomic discoveries into clinical tools provide valuable lessons to guide the translation of emerging insights from the genome into clinical end points that can affect the practice of cancer medicine.

A human 'cancer genome', or oncogenome, harbours numerous alterations at the level of the chromosomes, the chromatin (the fibres that constitute the chromosomes) and the nucleotides. These alterations include irreversible aberrations in the DNA sequence or structure and in the number of particular sequences, genes or chromosomes (that is, the copy number of the DNA). They also include potentially reversible changes, known as epigenetic modifications to the DNA and/or to the histone proteins, which are closely associated with the DNA in chromatin (Fig. 1). These reversible and irreversible changes can affect hundreds to thousands of genes and/or regulatory transcripts. Collectively, they result in the activation or inhibition of various biological events, thereby causing aspects of cancer pathophysiology, including angiogenesis, immune evasion, metastasis, and altered cell growth, death and metabolism[1].

Mining the cancer genome and epigenome for aberrations that control these processes has become a major activity in cancer research, because it is widely understood that these aberrations provide clues to the mechanisms of disease pathogenesis. These studies can inform efforts to identify molecular events that can be targeted for therapy and to discover molecular biomarkers (biological indicators) that aid in early detection, diagnosis, prognosis (that is, prediction of clinical outcome) and the prediction of responses to therapies. Recognizing this, many national and international efforts, including The Cancer Genome Atlas pilot project by the US National Cancer Institute and the National Human Genome Research Institute[2], have been initiated to accelerate the compilation of an atlas of alterations.

In recent years, cancer genomics — defined here as the study of the ensemble of DNA-associated abnormalities that allow and accompany cancer development — has exploded as a field, with studies facilitated by genome-wide, high-resolution, high-throughput platforms (Box 1). These technologies now yield informative, but dauntingly complex, multidimensional genomic data sets that describe in detail the myriad changes that occur within individual tumours and how these changes differ between individual tumours. Together with assays to detect these aberrations that are now used to stratify patients for treatment (discussed

later), these data sets are now transforming the practice of cancer medicine, as is shown by the success of therapies that target distinct molecular events resulting from genomic aberrations. For example, patients with mutations in the gene encoding the epidermal growth-factor receptor (EGFR) can be treated with gefitinib or erlotinib[3–5]; those with the *BCR–ABL* translocation, with imatinib mesylate[6]; and those with amplification (that is, increased copy number) of the oncogene *ERBB2* (also known as HER2 or NEU), with trastuzumab or lapatinib[7]. In parallel, assays for mutations in germline DNA can identify individuals who are at high risk of developing cancer. For example, mutations in *TP53* (which encodes the tumour-suppressor protein p53) are associated with Li–Fraumeni syndrome[8]; mutations in *BRCA1* or *BRCA2* indicate an increased risk of breast and ovarian cancer[9–11]; mutations in genes whose products are involved in DNA-mismatch repair (such as *MLH1*, *MSH2* or *MSH6*) are associated with hereditary non-polyposis colorectal cancer[12]; and mutations in *CDKN2A* (which encodes a tumour-suppressor protein known as INK4A (or p16), which is involved in regulating the cell cycle) indicate an increased risk of familial atypical multiple mole melanoma-pancreatic cancer[13].

These examples have demonstrated the promise of cancer genomics, stimulated rapid advances in genomic technologies and computational science, and galvanized an entire generation of multidisciplinary scientists to identify the next set of key therapeutic targets and disease biomarkers for cancer. Although there has been tremendous success in the rapid accumulation of genomic data, most of these enormous data sets have not yet been translated into meaningful clinical end points. In the past, the translation of each genomic aberration into improved management of patients has taken at least a decade and sometimes billions of dollars. Given this situation, it is important to understand the barriers that prevent more rapid and less costly conversion of genomic information into useful diagnostic tests and effective therapeutic agents. Is statistical significance in the absence of mechanistic insight sufficient to harness the full potential of these complex genomic data sets in a cost-efficient and effective way? Or is some degree of understanding of the molecular biological function required for efficient translation? The *BCR–ABL*, *ERBB2*

---

[1]Dana-Farber Cancer Institute and Harvard Medical School, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Lawrence Berkeley National Laboratory, Life Sciences Division, 1 Cyclotron Road, 977R225A, Berkeley 94720, California, USA.

and *EGFR* examples seem to support the view that coupling insights into the genome with pathobiological findings holds the greatest promise for making an impact in the clinic. In this article, we review lessons from past genomic discoveries that have been translated successfully into the clinic and describe strategies (including integrative analyses and model systems) that have been useful for the identification of genetic elements of interest (GEOIs). We conclude with a discussion of the challenges that are faced and potential ways to move forwards in this field.

Lessons from the past

There are several pioneering examples of genomic aberrations being discovered in cancer cells and the findings being successfully translated into therapeutic agents and tests for cancer risk, prognosis or response to therapy, with considerable effects on the practice of cancer medicine.

Although many of these successes predated the current genome-wide, high-throughput technologies — indeed, some resulted from decades of painstaking work — they nevertheless presage the translation of information from the cancer genome into clinical tools. These translational efforts can be considered in terms of the type of genomic aberration studied — translocations, gene amplification, mutations and germline susceptibility — and the examples described in this section might help to guide and accelerate translation of the genomic aberrations now being discovered.

Translocations

The first genomic aberration found to be associated consistently with a human malignancy (that is, recurrent) was the Philadelphia chromosome, discovered by Peter Nowell and David Hungerford in 1960 (discussed in
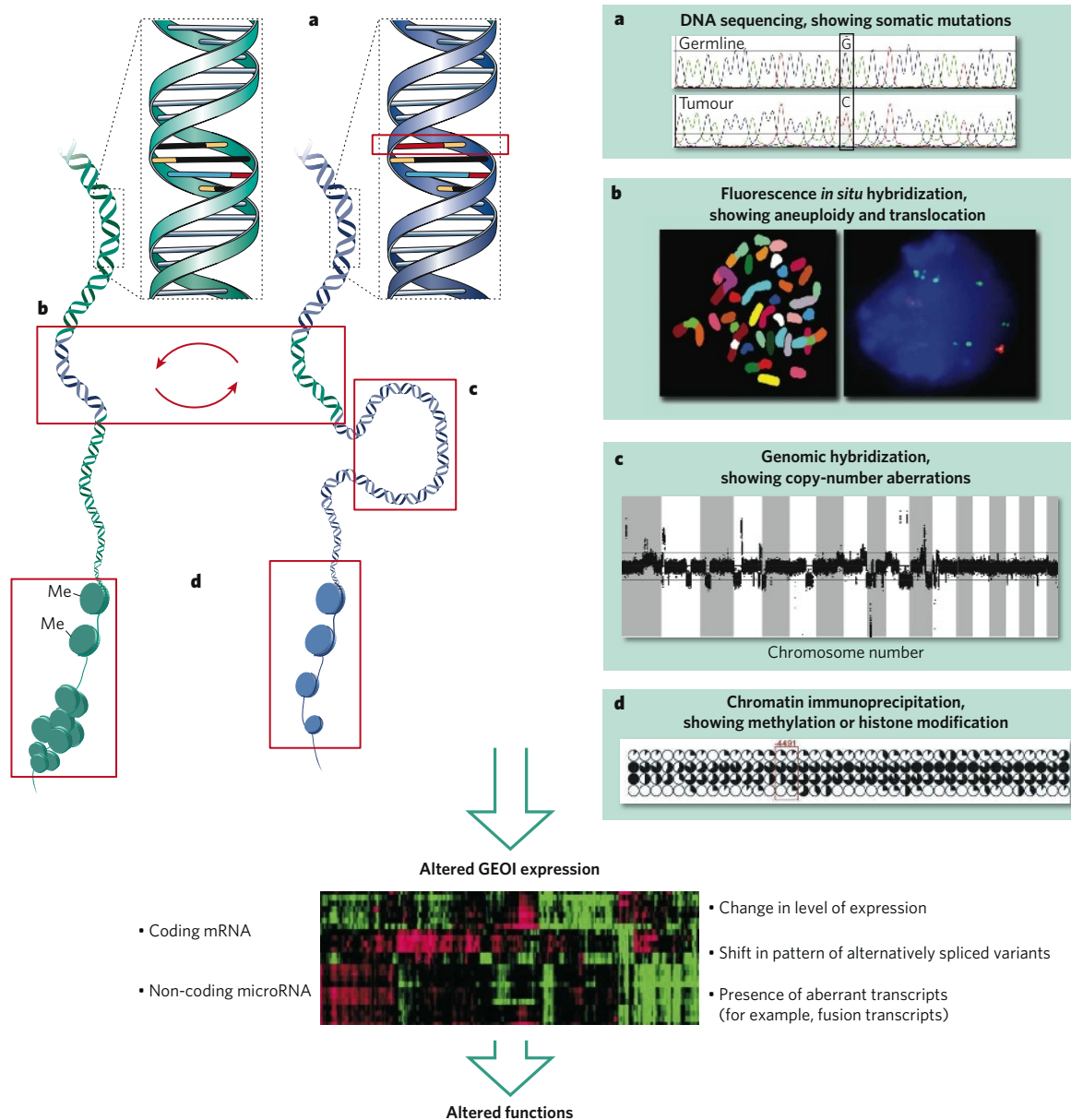
Figure 1 | Various types of genomic and epigenomic aberration in cancers. The main types of genomic and epigenomic aberration are illustrated together with examples of how they can be detected. **a**, Changes in DNA sequence, such as point mutations, can be assessed by DNA-sequencing techniques. **b**, Changes in genomic organization can be assessed by using fluorescence *in situ* hybridization. In the example shown, DNA segments are exchanged between the two (blue and green) DNA molecules. **c**, Changes in DNA copy number, such as those that result from amplification, can be assessed by using comparative genomic hybridization. **d**, Changes in DNA methylation and the resultant changes in chromatin structure can be assessed by using chromatin immunoprecipitation plus microarray analysis of immunoprecipitated DNA. Each of these types of change can alter the expression levels of genes or non-coding microRNAs (referred to here as genetic elements of interest, GEOIs), alter the splicing patterns of transcripts, or change gene function through mutation or through creating chimaeric genes. Many of these events can be as assessed by microarray analysis. These changes ultimately translate into altered functions, leading to a diseased state, such as cancer.

Box 1 | Techniques for analysing the cancer genome

Comprehensive analyses of the genome of various types of cancer cell — in terms of DNA copy number, DNA sequence, DNA organization, gene expression and epigenomic modification — are underway worldwide. A rapidly evolving suite of technological solutions is allowing cancer genomes to be characterized with remarkable resolution and accuracy. Several of the techniques used to analyse the various aberrations and modifications are summarized here.

Copy-number aberrations

Changes in the copy number of genetic regions or chromosomes across the entire genome of a cancer cell can be mapped onto a representation of the normal genome by using comparative genomic hybridization. This technique readily allows the genes involved in copy-number aberrations to be identified[31]. Modern analysis platforms for comparative genomic hybridization map copy-number changes onto DNA sequences arranged in microarrays[77] and allow these changes to be assessed quantitatively (including for individual alleles in some platforms) with sub-gene resolution. Even at this resolution, aberrations can be missed, especially when using platforms that are gene-oriented. Emerging next-generation technologies that efficiently sequence small genome fragments that have been collected randomly from tumour-cell genomes will complement such DNA-microarray-based strategies for analysing copy number. These work by sequencing tens of millions of short DNA fragments and then summing the number of fragments in equal-sized bins distributed along the genome. The relative number of DNA fragments in each bin is an estimate of the relative copy number at that genomic location. The resolution of this approach can be made arbitrarily high by sequencing to an increasing depth.

Structural aberrations

Structural changes can involve segmental deletions or insertions, and translocations or more complex rearrangements (for example, those occurring during gene amplification or copy-number change). These changes can be uncovered by using cytogenetic techniques such as banding analysis or fluorescence *in situ* hybridization or by using DNA-sequence-based strategies such as end-sequence profiling. End-sequence profiling is an adaptation of whole-genome shotgun sequencing that allows structural aberrations to be detected[23]. DNA from a tumour is cloned into a large-insert vector, and the ends of the resultant clones are sequenced and then mapped onto the normal human DNA sequence. Paired ends that map farther apart than the maximum size tolerated by the cloning vector indicate the presence of a structural aberration. This approach has the advantage that clones containing aberrant DNA from gene fusions can be sequenced to identify the exact DNA sequence at the breakpoint. But it has the disadvantage that millions of tumour DNA clones must be maintained. Alternatively, cloning strategies known as paired-end sequencing, which retain only the ends of the cloned DNA fragment, can be used[78]. These paired ends are then sequenced to identify structural aberrations (as described above). This strategy is efficient but does not yield the DNA sequence across the breakpoints.

DNA-sequence abnormalities

Recent efforts in large-scale DNA-sequence analysis have identified several hundred candidate genes that might have functional roles in various human cancers[40,41]. Some occur at a relatively high frequency, but most are present in only a few per cent of tumours. Results from the extensive sequencing and mutation-validation efforts that are now underway will be necessary to establish the prevalence of, and clinicopathological associations for, these genetic elements of interest (GEOIs). Both established and next-generation sequencing technologies will be brought to bear on this issue. There are several current techniques for DNA sequencing, including sequencing by hybridization, dideoxy sequencing and cyclic array sequencing. Sequencing by hybridization[79] is a DNA-array-based strategy in which mutations are detected based on the intensity of hybridization of sample DNA to microarrays comprising short oligonucleotide probes that are designed to be perfectly complementary to the reference sequence plus oligonucleotide probes that differ by one base at each 'substitution position' in the genome to be tested. This approach is well suited to resequencing. Dideoxy sequencing[80] is the current standard method for detecting mutations. It is typically applied to PCR products that result from the amplification of sample DNA by using primers that flank regions of interest, and it generates collections of DNA fragments in which each fragment terminates with a base-specific fluorescent label. The fragments are then separated according to size by using capillary electrophoresis, and the terminating base is identified by fluorescence emission analysis. Sequence 'reads' are generally about 750 bases. In most cases, dideoxy sequencing will not detect mutations that are present in less than about 20% of the cells represented in the PCR-amplified population. Mutations that have been discovered so far are summarized in the Catalogue of Somatic Mutations In Cancer (http://www.sanger.ac.uk/genetics/CGP/cosmic). The efficiency of sequencing can be increased by using matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry to measure the masses of DNA fragments generated by primer extension with dideoxy termination. However, sequence reads are typically less than 100 bases with this read-out method. Cyclic array sequencing allows millions to billions of DNA fragments to be sequenced in parallel by arranging these fragments on a sequencing substrate and using a cyclic enzymatic process to interrogate the sequence of all fragments in parallel[81,82]. Current read lengths range from about 30 bases to 300 bases, and the number of reads per analysis ranges from 0.3 million to 30 million. Cyclic array sequencing techniques facilitate the detection of rare mutations. Recent affinity-enrichment techniques allow subsets of the genome to be enriched before sequencing (for example, all known exons), thereby decreasing the cost of targeted sequencing[83].

Epigenomic analysis

It is clear that epigenomic modifications are major contributors to the formation and progression of tumours, especially during the early stages of tumour development. Several techniques for the genome-wide assessment of DNA methylation and chromatin structure have now been established, and others are emerging; these techniques help to elucidate the role of epigenomic modifications in the cancer genome. The five established techniques are restriction-landmark genomic scanning, microarray-based epigenomic analysis, reduced representation bisulphate sequencing, methylation-specific digital karyotyping, and chromatin immunoprecipitation plus microarray analysis. First, restriction-landmark genomic scanning[84] using methylation-sensitive enzymes was the first method developed as a genome-wide screen for methylation of CpG islands. This technique, which involves two restriction digests followed by electrophoretic separation, allows methylation to be analysed in up to 4,000 loci[85,86]. Second, microarray-based epigenomic analysis methods[87] involve hybridization of tumour and reference DNA samples to DNA microarrays. These microarrays comprise oligonucleotides derived from CpG-island sequences (which are generated by digestion of CpG islands with methylation-sensitive restriction enzymes that cleave preferentially within the islands). Comparing the signal intensities from the tumour and reference samples provides a profile of sequences that are methylated in the tumour but not in the references (or vice versa). Third, reduced representation bisulphate sequencing[88] is a genome-wide shotgun sequencing approach in which the tumour and reference DNA samples are treated with sodium bisulphate to convert cytosine to uracil while leaving 5-methylcytosine unconverted, then digested with a methylation-specific enzyme and sequenced. Comparison of CpG sequences in the tumour and reference genomes then reveals bisulphate-induced changes. This method is well suited to next-generation single-molecule sequencing strategies. Fourth, methylation-specific digital karyotyping[89] is a modified technique for DNA copy-number profiling[90]. Sequencing is carried out to accurately count tags to compare CpG sequences in tumour and reference samples, thereby allowing quantitative measurement of methylation events. Fifth, chromatin immunoprecipitation plus microarray analysis (ChIP on chip)[91] involves an initial immunoprecipitation step, thereby enriching DNA sequences associated with histone modifications (for example, methylation or acetylation of histone H3) for which specific antibodies are available. Immunoprecipitated DNA sequences are then analysed by using DNA-microarray-based methods or single-molecule DNA-sequencing strategies.

ref. 14). In the ensuing decades, cytogenetic and molecular studies showed this to be a translocation between chromosomes 9 and 22, resulting in a fusion product, *BCR–ABL*. As a result of this fusion, the activity of the non-receptor tyrosine kinase ABL is dysregulated in patients with chronic myeloid leukaemia or with some forms of acute lymphoblastic leukaemia. More than 30 years after the discovery of the Philadelphia chromosome, a small-molecule inhibitor of ABL, imatinib mesylate, was developed as an effective therapeutic agent against the effects of the *BCR–ABL* translocation in patients with chronic myeloid leukaemia[6]. However, despite marked initial responses, this targeted therapy does not lead to a lasting cure, because resistant cancer cells emerge[15]. Genomic analyses of the resistant cells showed that point mutations were acquired (and sometimes amplified) that abrogated the inhibitory effects of the drug. This result guided the development of new small-molecule inhibitors to counter this resistance mechanism, culminating in the recent approval of nilotinib and dasatinib by the US Food and Drug Administration (FDA)[16]. These findings suggest that the development of anticancer drugs will occur by an iterative process in which genomic analyses are first used to guide the development of targeted therapies and associated predictive biomarkers, and then genomic studies of resistant cancer cells aid in the development of second-generation and third-generation inhibitors to counter the mechanisms of resistance that have arisen against the first-generation inhibitors. Banking tumour tissues from patients who are sensitive or resistant to drugs will be essential to support these studies. In many cases, this will require biopsy of metastatic lesions, a process that is not regularly carried out in clinical trials. Another lesson from the imatinib mesylate story is that genomic analyses can guide the use of small-molecule inhibitors that are effective against several targets. Imatinib mesylate, for example, also inhibits the receptor tyrosine kinase c-KIT. Following genomic analyses of gastrointestinal stromal tumours (GIST sarcomas)[17] and mucosal melanomas[18], which showed that both cancers harbour *c-KIT* mutations, imatinib mesylate has been used successfully to treat patients with GIST sarcomas or mucosal melanomas[17–19].

Since the pioneering discovery of the Philadelphia chromosome, numerous recurrent translocations that cause cancer have been discovered in human leukaemias and lymphomas by using molecular cytogenetic analyses[20]. However, finding causal translocations in solid tumours has been difficult, possibly reflecting the complex genomic profiles and heterogeneous nature of these malignancies. With the current ability to analyse the genome, together with the sophisticated analytical approaches available and the ever increasing amounts of genomic information, recurrent structural aberrations are now being discovered in solid tumours and might be more prevalent than previously thought. A notable discovery is the high frequency of translocations between *TMPRSS2* (which is upregulated in response to androgenic hormones) and the *ETS*-family genes *ERG*, *ETV1* and *ETV4* (which encode transcription factors) in human prostate cancer. Using a new integrative analytical methodology called cancer outlier profile analysis, which identifies associations between genomic and transcriptional abnormalities, Arul Chinnaiyan and colleagues[21] identified a family of common translocations that brings *ETS*-family genes under the control of TMPRSS2, in effect placing the expression of these genes under androgen-mediated regulation. Molecular assays for fusion events are now being developed and evaluated for use as early detection markers for prostate cancers[22]. It is hoped that applying similar computational approaches to emerging multidimensional data sets will allow the detection of other causal structural aberrations in solid tumours. And this is only the beginning. Next-generation sequencing technologies that allow the entire genomes of tumour cells to be sequenced will be particularly valuable for discovering fusion genes and other structural rearrangements. The promise of this approach is illustrated by the remarkable structural complexity found in cancer genomes by using end-sequence profiling[23], genomic-region sequencing[24] or genome-wide parallel paired-end sequencing[25] (Box 1).

Gene amplification

Another prominent success story involves the now well-established *ERBB2* oncogene. *ERBB2*, which is homologous to mouse *Erbb* and

the gene encoding tumour antigen p185, was initially identified as a transforming oncogene in NIH/3T3 cells[26] and was also found to be amplified in human breast-cancer cell lines[27–29]. Shortly after these findings, *ERBB2* amplification was found in ~30% of primary breast-cancer tumours, and this amplification was associated with a short survival time and short time to relapse[30]. On the basis of these observations, trastuzumab (a monoclonal antibody specific for the extracellular domain of ERBB2) was developed to treat breast tumours that had *ERBB2* amplification[7]. Clinical introduction of trastuzumab was guided by molecular assays for *ERBB2* amplification[31] or overexpression. More recently, molecular diagnostic assays that assess *ERBB2* amplification or overexpression have guided clinical use of the small molecule lapatinib, which targets ERBB2 and EGFR[32].

Mutations

Since the completion of the Human Genome Project, several important discoveries in genomics have come from the systematic resequencing of genes, gene families or genes in pathways that are relevant to cancer. One of the first, and perhaps most celebrated, successes from such large-scale resequencing projects was the discovery that *BRAF*, which encodes a serine/threonine kinase, frequently contains activating somatic mutations: in 60% of malignant melanomas, in 10% of colorectal cancers and in a smaller percentage of other cancers[33]. This discovery has driven many programmes aimed at developing BRAF inhibitors, and several drugs are now in clinical trials. Other notable discoveries from large-scale resequencing efforts include frequent mutations in *PIK3CA*[34] (which encodes the catalytic subunit of phosphatidylinositol-3-OH kinase) and *AKT1* (ref. 35) (which encodes a serine/threonine kinase) in many cancer types, as well as in *ERBB2* and *EGFR* in non-small-cell lung cancer[36,37]. In addition to gender, ethnicity, smoking history and the histopathological subtype of the cancer, it was found that the mutation status of *EGFR* predicts responses to treatment with the EGFR inhibitors gefitinib or erlotinib in patients with advanced non-small-cell lung cancer[3–5]. Testing for mutations in *EGFR* before decisions are made about treatment with EGFR inhibitors is becoming routine[37]. The ability to determine *EGFR* genotype retrospectively by using banked tumour tissues with matched germline DNA from the ongoing clinical trials was crucial for allowing the stratification of responders and for showing efficacy[38]. These studies highlight the importance of uniformly collecting pretreatment and post-treatment tumour specimens with matched normal controls from clinical trials.

Germline susceptibility

In addition to its impact on somatic genetics studies, genomics is revolutionizing the search for germline genes that confer susceptibility to cancer and for polymorphisms that are responsible for inherited predisposition to disease, including cancers. One of the early successes in this area was the discovery that inactivating mutations in *BRCA1* are associated with familial breast cancer[9,10]. Genetic screening for germline mutations in *BRCA1* — and now in a second cancer-susceptibility gene, *BRCA2* (ref. 11) — is being rolled out worldwide to identify patients who are at a high risk of developing early-onset breast and ovarian cancer. Moreover, the knowledge that BRCA1 is required for error-free repair of DNA double-strand breaks led to the development of inhibitors of poly(ADP-ribose) polymerase 1 (PARP1, an enzyme involved in the recognition of DNA single-strand breaks)[39]. These and subsequent studies showed that discovery of inactivating germline mutations associated with increased susceptibility to cancer can be guided by analyses of loss of heterozygosity or reduction in DNA copy number and/or DNA methylation in the tumours that eventually develop. Applying current (and future) genomic technologies in coordinated germline and tumour studies should considerably accelerate the discovery of susceptibility genes of this class, thereby increasing our ability to identify high-risk individuals who can then be managed using aggressive surveillance and prevention strategies. Identifying susceptibility genes by this method will require the coordinated collection of tumour specimens, together with germline DNA, in large-cohort genetic susceptibility studies.

Making sense of the cancer genome

Empowered by the improved ability to survey the cancer genome with increasing accuracy and resolution, numerous studies have been carried out or initiated with the hope of discovering the next EGFR, ERBB2 or BRAF. Instead, these analyses are uncovering hundreds of recurrent genomic or genetic alterations that affect thousands of GEOIs — including annotated genes, non-coding microRNAs and other conserved elements — that might contribute to the pathophysiology of human cancers. The nature and strength of each GEOI, the certainty of its contribution to cancer, and therefore its translational importance, varies substantially. Some GEOIs will be strong, causal 'drivers' of important cancer hallmarks[1]. Others will be weaker but important 'contributors' to the development of cancer pathophysiology. And many will be genomic 'noise' (or 'passengers'): that is, elements that are biologically 'neutral' and have been accumulated by chance during the cancer's lifespan. Distinguishing the drivers and contributors from the passengers is a central challenge in genomic research. This is made more difficult by the diversity of GEOI function and the likelihood that GEOI function might depend on the tumour type (or subtype), as well as on the tumour microenvironment.

The assignment of GEOIs as drivers is compelling in the case of high-frequency events: for example, the amplification of regions that contain EGFR in glioblastomas (in 45% of tumours) or ERBB2 in breast cancer (in 20% of tumours); deletions of regions that contain CDKN2A or the tumour-suppressor gene PTEN (in many solid tumours); or mutations in TP53, RAS, BRAF or PIK3CA in a wide range of solid tumours (see the Cancer Gene Census, http://www.sanger.ac.uk/genetics/CGP/Census). Such assignments rest on the weight of functional evidence built up over decades, a luxury not afforded for GEOIs that are being found and will be found by using modern high-throughput genomic technologies. Furthermore, these prominent 'gene mountains' seem to be few and far between relative to the numerous 'hills and valleys' stretching broadly over large regions of the cancer genome[40,41]. Which of these GEOIs are involved in the crucial paths to malignancy? And what are their relative contributions? These are challenging questions without simple answers, but progress can be made by integrating data from multiple systems and then searching for common patterns (Fig. 2): that is, searching for GEOIs that are recurrently dysregulated by multiple mechanisms in several biological systems. In this section, we discuss several approaches that have been used successfully to find drivers and contributors — the needles in the haystack of cancer genome data — including integrative analyses of multidimensional data, interspecies comparative genomics and analyses of human cancer cell-line systems.

Integrative analyses of multidimensional data

The cancer genome can be dysregulated through multiple mechanisms. These include modifications to the DNA and the histones, changes in the DNA structure and copy number, and mutations in the coding and non-coding sequences. These changes can lead to alterations in transcription, translation, post-translational modification and, ultimately, gene and protein function. Technological advances that allow the cancer genome to be examined in multiple 'omic' dimensions are helping to focus the search for drivers and contributors, by uncovering GEOIs that tend to be dysregulated by several mechanisms. A classic example is the tumour-suppressor protein INK4A (encoded by CDKN2A), which can be inactivated in three ways: through the homozygous deletion of 9p21 or the region of 9p21 that contains CDKN2A; through the epigenetic silencing of gene expression (by promoter methylation); or through point mutations that cripple the function of INK4A[42]. Similarly, the PIK3CA oncogene can be activated through amplification and overexpression[43] and/or through activating mutations[34]. Such dysregulation through multiple mechanisms is clearly illustrated when examining well-known oncogenes in a typical signalling pathway (Fig. 3). In other words, if a genetic element is important, then the cancer will find a way to dysregulate it by any means possible. For this reason, the targeted resequencing of genes located in regions of amplification has borne fruit, such as identifying the c-KIT oncogene as a therapeutic target for mucosal and acral melanomas[18]. Thus, data showing that a GEOI

can be dysregulated in several complementary ways in cancer, through the integration of more than one dimension of genomic information, provide strong evidence that a GEOI is likely to be pathogenetic. The current large-scale cancer genome projects that are carrying out genome-wide characterization in a coordinated and comprehensive manner will be the most powerful at leveraging such multidimensional data for integrative analyses. In addition, integration across tumour types can be highly informative, because it is clear that the mechanisms of dysregulation of many oncogenes, including MYC, EGFR, AKT1, RAS, TP53, PTEN and CDKN2A, vary according to tumour type. For example, genes, such as MYC, that are activated by translocation in leukaemias can be activated by amplification in solid tumours. The convergence of genomic data that implicate a particular GEOI across tumour types can help to rapidly prioritize GEOIs that are likely to have broad importance. As a by-product, it is probable that the power of genomic biomarkers to determine prognosis or predict responses to therapies will increase substantially if assays are developed to assess the cumulative effect of all mechanisms of dysregulation, including effects on protein structure and abundance.

Interspecies comparative cancer genomics

Another approach to uncovering drivers and contributors is to use evolutionary conservation as a guide. This can be a powerful way to find oncogenes, because genes that are involved in pathways that are dysregulated in cancers — such as receptor-tyrosine-kinase signalling, cell-cycle regulation and apoptosis — are strongly conserved across species[44,45]. This comparative approach was enormously helpful in refining the draft of the human genome sequence. With respect to cancer, it has been established that oncogenes from one species can induce the malignant transformation of cells from different species, despite poor sequence conservation (for example, the Drosophila spp. homologue of MYC, diminutive, can transform rodent cells[46]). Recent large-scale, cross-species comparisons have established that mouse and human tumours sustain orthologous genomic events in diverse tumour types[47–49]. This finding supports the view that genomic alterations conserved across species are more likely to represent crucial events in tumorigenesis and that using evolutionary conservation as a filter can provide a powerful solution to the central problem of noise in genomic data sets.

Early studies of cancer across species involved histopathological diagnoses, but such cross-species comparisons now include genetic and genomic analyses to show, for example, that genetically engineered mice can be used to model genetic aspects of human cancer. That mouse models are valid for studying human cancer is exemplified by cross-species conservation of gene-expression patterns that result in activation of the gene encoding Ki-RAS in lung cancers[50], as well as conservation of somatic mutations in the gene encoding NOTCH1 in mouse and human T-cell acute lymphoblastic leukaemia[51]. These findings were followed by studies providing proof of the concept that comparing genomic profiles of mouse and human tumours allows previously unidentified oncogenes to be uncovered[47,48]. In one of these studies, by Minjung Kim et al.[47], the ability to manipulate stages of mouse tumour development in vivo — from regression to recurrence to escape — was used to force the selection of aberrations conferring metastatic capability on tumours. Genome-wide copy-number profiles of these 'escaper' tumours revealed focal amplification in regions syntenic (that is, on a chromosomal region of common evolutionary ancestry) to human 6p24–25, a region that sustains copy-number gain in 36% of metastatic melanomas but not in primary melanomas[52]. Although 6p gain is highly recurrent, indicative of potential pathogenetic and/or prognostic importance in human tumours, the large region of amplification in human tumours makes the identification of drivers and contributors difficult or even impossible. Given the focal nature of the event in mice, cross-species comparison was able to narrow down the area of interest to an 850-kilobase region encompassing only eight annotated genes, with NEDD9 (which encodes an adaptor protein) as a putative driver. With that information as a guide, further functional and clinicopathological studies documented the metastasis-promoting activities of
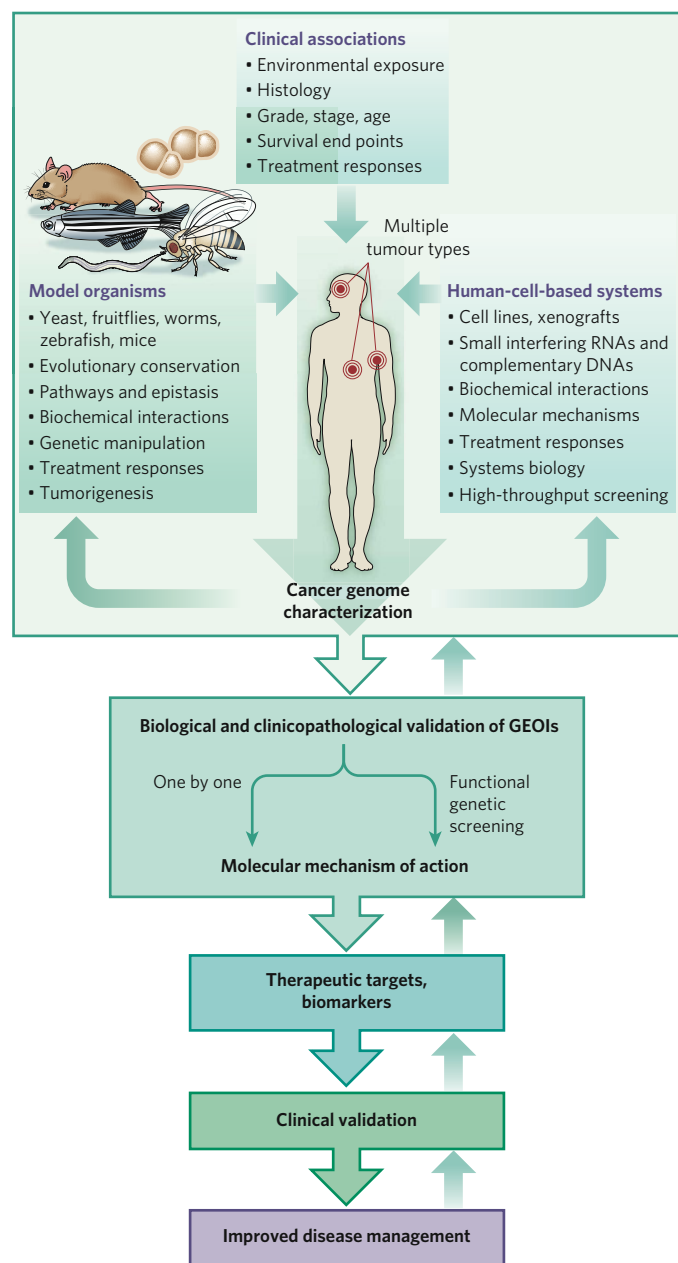
Figure 2 | Integration of complex multidimensional genomic data with insights from other model systems. The identification of cancer drivers or contributors from multidimensional genomic data (such as that shown in Fig. 1) from a particular human tumour type can be facilitated by integration with similar data from other tumour types (for example, by searching for GEOIs that are common to cancers of different lineages). Incorporating clinical information into the analysis of this genomic data helps to narrow the focus to clinically relevant GEOIs. This genomic knowledge can be further filtered by integrating it with insights obtained from studies in model systems. These systems can include non-human model organisms (such as mice, zebrafish, nematodes, fruitflies and yeast), which can be studied to identify evolutionarily conserved GEOIs, to define pathways that GEOIs influence and to elucidate the roles of GEOIs in normal development. Cell-line model systems also can be useful, particularly for functional genetic screening or monitoring responses to drugs. This type of integrative analysis, which extends beyond the cancer genome, is an informative way to identify GEOIs that are likely to be drivers or contributors. After such GEOIs have been identified, they need to undergo stringent biological and clinicopathological validation (Box 2), a labour-intensive process that can be accelerated by carrying out functional screening with a library of GEOIs rather than by assessing one GEOI at a time. For successful translation into the clinic — that is, development of a therapeutic agent that targets the GEOI or a biomarker for the GEOI — a basic understanding of the molecular mechanism of action of the GEOI is helpful, particularly in terms of the specific cellular and genetic context in which it maintains the tumour. Such a biomarker or therapeutic agent will then need to be clinically validated before it can be adopted for routine clinical practice. At each step of this process, the results can be fed back to inform and refine the analyses and to help improve the validation platforms.

NEDD9 and uncovered its molecular mechanism of action (interaction with focal adhesion kinase). Likewise, when looking at recurrent copy-number aberrations in tumours with *ERBB2* amplification, comparisons between human breast tumours and a transgenic mouse model (in which an oncogenic form of *Erbb2* called *NeuNT* was expressed under the control of the endogenous *Erbb2* promoter) implicated the genes encoding GRB7 and 14-3-3-σ as contributors to the ERBB2-mediated oncogenic process[53].

Although syntenic aberrations have been observed between mouse and human tumours, it is important to note that the genomes of most mouse tumours accumulate far fewer aberrations than do solid tumours in humans. For example, in oncogene-driven mouse models of cancer, tumours often have few or no copy-number aberrations, and the infrequent (and typically simple) copy-number aberrations that are present presumably occur only under strong selective pressures. This simplicity facilitates the identification of drivers and contributors targeted by such copy-number aberrations, as exemplified by the studies of Kim *et al.*[47] (discussed earlier) and Lars Zender *et al.*[48]. The disadvantage, however, is that this method does not lend itself to widespread use of cross-species comparison.

On the basis of observations that DNA-breakage events induced by telomere dysfunction can drive regional amplifications and deletions and that laboratory mice do not experience telomere-based crisis, Ronald DePinho and colleagues knocked out the gene encoding the RNA component of the telomerase holoenzyme from the mouse germ line in an effort to humanize the mouse genome. The resultant telomerase-deficient mice experienced progressive shortening of telomeres with each successive generation of mice, eventually leading to telomere-based crisis[54]. Tumours from these animals indeed showed high levels of instability, harbouring numerous non-reciprocal translocations and complex copy-number aberrations[55–57]. A genome-wide comparison of such genome-unstable mouse tumours with several human cancers of diverse origins showed non-random overlaps between the copy-number aberrations. This finding proves that mouse and human tumours experience common biological processes that are driven by orthologous genetic events[49].

Attesting to the potential of such cross-species comparisons in oncogene discovery, the focused resequencing of GEOIs within syntenic deletions uncovered a high frequency of mutations in *FBXW7* and *PTEN* in human T-cell acute lymphoblastic leukaemia[49]. Mutations in *PTEN* were

also shown to modify responses to NOTCH1 inhibitors in the clinic[58]. These studies support the idea that cross-species synteny is both a measure of validation, by virtue of evolutionary conservation and use of different genetic mechanisms (that is, a GEOI can be dysregulated by different mechanisms, such as mutation and copy number), and a guide for discriminating drivers and contributors from passengers.

Another way in which mice are valuable for comparative genomic studies is in the identification of susceptibility loci. Extending the concepts used to identify *BRCA1*, it might be expected that mutations or polymorphisms that contribute to cancer susceptibility are subjected to positive selection during the 'evolution' of the cancer genome. Thus, these mutations or polymorphism might be found by allele-specific analysis of copy number and gene expression in defined model systems. For example, using genomic strategies, Allan Balmain and colleagues[59,60] identified that polymorphic variants of *AURKA* (also known as *STK15*), which encodes an aurora kinase, are associated with an increased risk of developing cancer at several sites in humans. These studies began by analysing the position of quantitative trait loci that control susceptibility to skin-tumour formation in mice from interspecific crosses (*Mus musculus* × *Mus spretus*). One of these loci, *Skts13*, was orthologous to a region that is frequently increased in copy number in human cancers of the breast, colon and ovary; this region, 20q13, contains the gene encoding AURKA. Analyses of the expression of the mouse orthologue of *AURKA*, *Stk6*, showed an allele-specific difference in the mouse interspecific crosses, and copy-number analyses of two alleles *AURKA* 91A and *AURKA* 91T showed that *AURKA* 91A is preferentially amplified in human colon tumours. A subsequent meta-analysis of the association between the alleles *AURKA* 91T and *AURKA* 91A and the risk of developing cancer of the colon, breast, prostate, skin, lung and oesophagus showed an increased risk in both homozygotes and heterozygotes. These results confirmed that the *AURKA* 91A is a low-penetrance cancer-susceptibility allele that increases the risk of developing many cancer types. This integrative analysis of quantitative cancer traits in mice, allele-specific copy-number change and expression, and susceptibility to cancer in large population-based studies serves as a model for the definitive identification of the (probably large number of) low-penetrance, high-prevalence polymorphisms that influence cancer risk.

Finally, model systems, including mouse models, are well suited to forward genetic screens, in which researchers can 'listen' and let the cancer cells 'tell' which events are required or preferred on the path towards full malignant transformation. For example, retroviral insertional mutagenesis in mice has yielded recurrent and common insertion sites at loci containing genes such as *Ras*, *Myc*, *Notch1*, *Flt3*, *c-Kit* and *Tp53* (ref. 61), attesting to the power of this method to identify oncogenes when the results are integrated with existing and emerging human cancer genome data.

Cell-line model systems

Much of our understanding of tumour cell biology, including aspects of gene regulation and signalling, has come from studies of tumour cells in culture. The roughly 50,000 publications describing uses of the HeLa cell line and the 20,000 publications describing uses of the NIH/3T3 cell line attest to this fact. That said, established tumour cell lines grown on plastic dishes, in three-dimensional cultures or in immunocompromised mice cannot fully recapitulate all the biological aspects of tumours that are growing in the complex human microenvironment. Nor can any model fully represent the responses of the various human tumours to therapy — in part because of differences in the biological environment and in part because the models do not capture the range of biological, genomic and epigenomic diversity found in human tumours. Therefore, it is expected that each model system has strengths and weaknesses. Mice are one such system. As we have described, and as is discussed in greater detail elsewhere[62], the value of mouse models is unequivocal. As long as researchers are aware of the limitations of any one model, then the information that such a system offers can be used. Integrating data from several models will help to build a true picture of cancer.

So, what can be learned about genomic aberrations by studying cell-line models? And why are these models important? To put it simply, cell
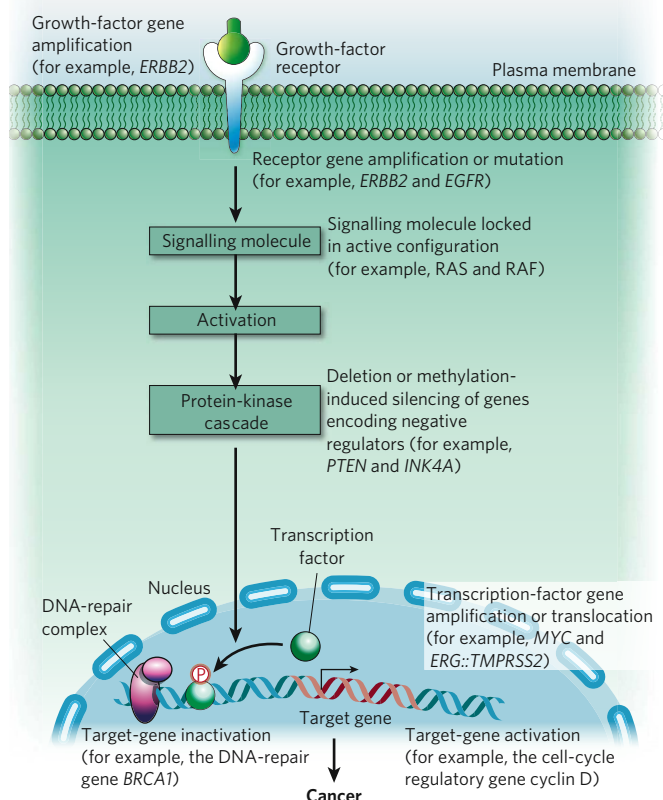

Figure 3 | Disruption of intracellular signalling by alterations in the cancer genome. A simplified signalling pathway is depicted to highlight known examples of *bona fide* oncogenes that are subjected to dysregulation by various mechanisms. It is clear that a signalling pathway can be disrupted at multiple points, and a variety of genomic and epigenomic alterations can contribute to this, ultimately leading to cancer.

lines are essential for the functional and biological validation of GEOIs (Box 2). Almost without exception, the functional validation of a GEOI and establishment of its molecular basis of action begins with various cell-line model systems, including established tumour cell lines (which are versatile and easy to manipulate). These systems allow the possible roles of GEOIs in the pathophysiology of cancer to be tested. For example, the driver or contributor role of a GEOI found in a region of recurrent amplification might be studied by assessing the consequences of enforced expression of the GEOI in cell lines in which it is expressed at a normal level. Likewise, the role of a GEOI in a region of recurrent deletion might be assessed by decreasing its expression by using RNA interference (RNAi)-mediated knockdown in a cell line in which it is expressed at a normal level. Cell lines derived from tumours in which GEOIs are dysregulated by genomic or epigenomic aberrations are valuable 'experiments of nature' that also provide information about GEOI function, for example through assessing the biological consequences of restoring dysregulated GEOI expression to levels that are closer to normal.

A major obstacle to the accurate interpretation of functional data derived from established tumour cell lines is the lack of clarity about the complements of genetic alterations that these cell lines carry. It has become clear that the genotype of the system — be it a cell line, a model or even a patient — can dictate the behaviour of tumour cells and can alter their response to a manipulation such as RNAi-mediated knockdown or pharmacological inhibition. As is the case for the original tumours from which they were derived, no two tumour cell lines are alike. Moreover, there is the legitimate concern that genomic aberrations will be gained or lost during extended passages in culture. Therefore, it is important that cell-line models — whether grown on plastic, in three-dimensional culture or in xenografts (that is, a grafted into a different species) — are

Box 2 | Validation of GEOIs

The points below outline the basic approach to validating GEOIs that have been identified in the cancer genome. GEOIs need to be validated in terms of their biological activity and their clinicopathological association, and each validation should be confirmed by using several assays. It is important to note that it is the cumulative weight of evidence — as assessed by several of the assays outlined below but not any single assay — that determines whether a GEOI contributes to, or drives, cancer.

Biological validation
The types of assay for biological validation are listed.

Model systems for manipulation of GEOIs
• Established mouse or human tumour cell lines with detailed genomic characterization
• Relatively naive, non-transformed primary cell cultures established from normal tissues
• Genetically engineered animal models for *in vivo* studies

Candidate GEOI manipulation
• Loss of function, by RNAi-mediated knockdown of the GEOI or by pharmacological inhibition using available drugs
• Gain of function, by expression of cDNA containing the GEOI or by pharmacological activation using available drugs

Functional assays for biological activity
• Cell proliferation and/or apoptosis, and migration and invasion in two-dimensional or three-dimensional culture models
• Anchorage independence *in vitro*
• Migration, invasion and *in vivo* lung seeding through tail-vein injection
• Tumorigenicity of subcutaneous or orthotopic xenotransplants

GEOI-specific assays
• Biochemical or intracellular signalling activities

Clinicopathological validation
The properties of GEOIs that are likely to drive or contribute to cancer are listed, together with ways to search for these properties.

Evidence of dysregulation at the DNA level through various mechanisms
• Search for mutations and copy-number changes (Box 1)
• Search for epigenetic modifications (Box 1)

Prevalence
• Assay genetic and/or epigenetic events in a large number of tumours across a broad range of cancer types

Evidence of altered expression
• Measure level of expression
• Search for altered splicing variants
• Search for novel transcripts of translocated regions
• Assess changes in proteins

Correlation with clinical parameters
• Cancer subtype
• Survival duration or response to treatment

specimens have led to the identification of the transcription factor REST as a tumour-suppressor protein in colon cancer[63], *IKBKE* (which encodes a signalling molecule) as an oncogene in breast cancers[64], and *PIK3CA* mutations as important determinants of resistance to treatment with trastuzumab[65].

Cell lines are also important models for assessing drug sensitivity and resistance in the quest to identify biomarkers that can guide early-phase clinical-trial studies; to identify drugs that might be effective in cancer subtypes that are resistant to the drug(s) used in the current standard of care; and to identify effective drug combinations. Although still in its infancy, an increasing number of studies support the concept that molecular biomarkers for predicting drug responses can be uncovered by analysing how molecularly characterized tumour cell lines respond to particular chemotherapeutic agents (which target molecular mechanisms that are intrinsic to the tumour cells)[66–70]. As a corollary, these analyses also identify drugs with a high specificity for subsets of tumour cells defined by certain molecular characteristics. Examples are *in vitro* analyses that predict the known sensitivities to trastuzumab[71] and lapatinib[68] of tumours in which ERBB2 has been amplified, sensitivity to gefitinib of tumours harbouring *EGFR* mutations[3–5], resistance to gefitinib conferred by an acquired mutation in *EGFR*[72], and resistance to imatinib mesylate in tumours with mutated or amplified *BCR–ABL*[73].

Using cell-line model systems that include large numbers of independent, established tumour cell lines of broad molecular and cellular diversity, together with comprehensive genomic characterization, can be and will be tremendously effective for translating genomic insights into clinical end points. But these systems could be further improved by developing co-culture or three-dimensional culture conditions that more closely model *in vivo* microenvironments, as well as by developing strategies to establish primary or short-term cultures that minimize the 'culture shock' associated with adapting to plastics.

A molecular understanding

The identification of driver or contributor GEOIs, especially the weaker or less prevalent ones, can be greatly accelerated by integrative analyses of multidimensional data and by comparisons with data from multiple model systems or species (Fig. 2). But identification of a GEOI is insufficient for its translation into a clinical end point. Cancer is a complex and heterogeneous collection of disease entities that are defined by clinical, histopathological and genetic parameters. Given this disease heterogeneity, even if a strong correlation between a GEOI and cancer is found in the laboratory in a test validation set (for example, a collection of genomic data, behaviour in a model system or even responses in a clinical trial), this correlation, no matter how significant, might not apply to every patient or trial subject. Without a definition of the genomic and biological context under which a GEOI exerts its cancer-associated activities, the full diagnostic and prognostic and therapeutic value of these genomic insights will not be realized.

Consider the example of *EGFR* mutations in non-small-cell lung cancer and glioblastoma multiforme (GBM). Mutational activation of *EGFR* in non-small-cell lung cancer is present in a subpopulation of patients who are highly responsive to targeted inhibition of EGFR. The proportion of patients with non-small-cell lung cancer who have an activating mutation in *EGFR* is small (about 10% in studies carried out in the United States and somewhat higher in Asian populations)[37]. Thus, the response of these patients to gefitinib, which inhibits the tyrosine-kinase activity of EGFR, would not have emerged in the absence of genetic stratification of this clinically distinct population. Conversely, amplification of a mutant form of EGFR known as EGFRvIII is prevalent in GBM (in about 45% of primary GBM cases)[74], yet EGFR-specific tyrosine-kinase inhibitors have strikingly little clinical effect. A positive, albeit transient, clinical response has been detected in subsets of patients in whom *EGFR* is amplified or mutated but *PTEN* is intact[75], indicating that this key molecule downstream of EGFR in the signalling pathway can modify the biological response of the tumour (Fig. 3). However, these positive responses do not last, despite documented pharmacological extinction (that is, inactivation) of mutated or amplified *EGFR*. In

subjected to the same level of comprehensive genomic characterization as human tumour specimens. In this way, the interpretation of functional studies can be guided by the knowledge of the similarities and differences between the cell lines and tumours that they are intended to model. It is also important that any cell-line system used for functional studies of the cancer genome comprises multiple independent cell lines that are molecularly diverse. If there is sufficient diversity, analyses of such cell-line collections minimize the risk that the elucidated function of an aberration will be idiosyncratic to a particular cell line.

As is the case for model organisms, forward genetic screening using a tumour cell-line model (particularly given recent advances in RNAi technology) can be used to identify cancer-relevant genes. Such *in vitro* screens are limited by the kinds of phenotypes amenable to high-throughput screens in culture (such as viability and growth assays). Nonetheless, recent studies that combine high-throughput RNAi-based screening with *in vitro* genomic profiling of primary human tumour

this case, the proteomic profiling of receptor-tyrosine-kinase activation patterns in solid tumours, including GBM and lung cancer, has provided a rational explanation for the patterns of clinical responses. Specifically, Jayne Stommel et al.[76] showed that established GBM cell lines, GBM xenotransplants and GBM primary tumour specimens from patients contain several coactivated receptor tyrosine kinases and that inhibition of EGFR alone can lead to its replacement with other coactivated receptor tyrosine kinases in the phosphatidylinositol-3-OH-kinase (PI(3)K) signalling complex, thus maintaining downstream signalling and cell survival. Signalling downstream of PI(3)K was extinguished only when multiple receptor tyrosine kinases were targeted by RNAi or by a combination receptor-tyrosine-kinase inhibitor[76]. Thus, the integration of genomic and proteomic insights with the molecular dissection of the signalling complex now provides a more accurate blueprint for the rational deployment of receptor-tyrosine-kinase inhibitors for treating GBM, tumours of the lung and other solid tumours.

Establishing the molecular basis of action of a GEOI in a specific tumour-biological context is perhaps the most difficult step in cancer genomics. Compounding the challenges of lengthy and laborious functional and clinicopathological validation (Box 2) is the biological phenomenon of false negatives. False negatives can arise in many ways; for example, when the cancer-associated biological activities of a GEOI (such as interaction with the host stroma) are not captured by standard cell-based assays; when a GEOI has a relevant role but only in a particular cellular or genetic context that is not recreated in the validation assay; and when a GEOI contributes only part of the overall activity conferred by a genomic event (so that the activity of a single GEOI is negligible in the absence of this cooperating partner or partners). Therefore, validation must not rely on just a single type of assay that involves a single manipulation.

Gain-of-function and loss-of-function manipulations for multiple tumour phenotypes using multiple cell lines should be carried out to search for the context in which biological activity can be uncovered. This process can be aided by knowledge obtained from other analyses, such as information about the biology of the tumour, the gene family of the GEOI, the pathways that the GEOI product is involved in, and insights from integrative analyses that nominated the GEOI. For example, if a GEOI identified by integrative genomic analyses is prioritized further on the basis of its known role in neural stem-cell homeostasis, then the next step would be to assess how manipulation of the GEOI affects the renewal, maintenance and differentiation of neural stem cells, in addition to carrying out the more generic assays of anchorage independence or cell proliferation (Box 2). Similarly, if a GEOI is identified in a subset of tumours with a particular genotype (such as with activated RAS or a mutation in EGFR), then its biological importance needs to be assayed in the appropriate context. This process has been demonstrated in two recent studies[47,48]. Kim et al.[47] showed that NEDD9 had gain-of-function pro-invasion activities only in cells in which BRAF or RAS was concomitantly activated, an experimental design that was informed by the characteristics of the metastatic escapers harbouring NEDD9 amplification. Zender et al.[48] showed that the inhibitor of apoptosis IAP1 (also known as BIRC2) and the transcription factor YAP had oncogenic activities in Tp53[+/−] hepatoblasts with Myc activation but not in those with Akt1 or Ras activation. This finding is consistent with the presence of an amplicon in the chromosomal region 9qA1 (which contains the genes encoding IAP1 and YAP) in this mouse model of hepatocellular carcinoma. In the study by Zender et al.[48], both IAP1 and YAP were shown to be targets of 9qA1 amplification, showing that a single genomic aberration can dysregulate more than one gene that contributes to the pathophysiology of the cancer. The chances of missing important GEOIs in a region of recurrent aberration can be reduced by using efficient functional genomic assays to assess the consequences of changing the expression levels of all GEOIs associated with the aberration. For example, genetic screens can be carried out with low-complexity libraries representing GEOIs resident in a particular genomic event (which is especially useful for regions that are large and gene-rich), allowing the identification of cooperating contributors (which together confer the

biological advantage selected for in the cancer cells). This functional genomic approach will be important for sorting out which of the less impressive 'hills and valleys' are biologically important.

Similarly challenging is the issue of biological false positives. For example, an RNAi-mediated loss-of-function assay is a powerful way to determine whether the expression of a GEOI is required in a cell for a specific tumorigenic phenotype (such as cell survival, anchorage independence or invasion). However, given the innumerable genetic and epigenetic alterations that are present in established tumour cells (and, consequently, the altered signalling between pathways and networks), the observed phenotype might be an artefact. In this case, finding a complementary gain-of-function activity can help to increase the evidence in support of a particular GEOI being a true driver or contributor to cancer. In addition, the type of functional activity also conveys a different level of confidence; for example, anchorage-independent growth in soft agar is a more stringent assay than increased proliferation in fully supplemented culture medium.

Biological false positives can also emerge as a direct consequence of the artificial nature of the assays used. Consider the possibility that overexpression of a GEOI confers a strong anchorage-independent phenotype; this effect might, however, result from the supraphysiological level of expression in vitro. Conversely, knockdown of a GEOI might result in cell death because its expression is required for the survival of all cells not just cancerous ones. To this end, clinicopathological validation through analysis of the DNA, messenger RNA and protein levels in normal samples and tumour samples arranged in microarrays can provide support for cancer relevance, by demonstrating the prevalence of genomic aberrations or dysregulated GEOI expression in large independent cohorts of specific tumour types. This can be particularly informative if the tumour cohorts are annotated with the clinical outcome because such a survey will not only add to the evidence but also provide invaluable insight into possible clinical contexts for therapeutic development. Ultimately, it is the cumulative weight of evidence based on the strength of particular functional activities, the magnitude of clinicopathological data and the importance of mechanistic clues that provides the confidence to assign a GEOI as a cancer-relevant driver or contributor rather than a mere passenger.

Conclusion

Cancer is the phenotypic end point of numerous genomic and/or epigenomic alterations that have accumulated within cells, and of the interactions of such altered cells with the stromal components in a unique host microenvironment. Some of the major challenges in translating the knowledge gained from cancer genomics into clinical practice stem from the fact that many cancer-associated changes in the genome are noise, as well as from the incomplete understanding of the biological functions of many of the genetic elements that are present in recurrent genomic alterations. Compounding these issues is the unfortunate reality that cancer is a highly complex, nimble and versatile disease.

We argue here that making sense of this complexity can be greatly facilitated by integrating genomic and biological insights from model systems with clinical knowledge of the disease. Translation can further be accelerated by rigorous biological validation and mechanistic exploration in preclinical settings to better define the clinical context(s) in which a genetic element (or components of the pathways or networks that is involved in) is an effective point of intervention for therapy. At the same time, we need to consider that the current understanding of what makes a strong driver, a cooperating contributor or, for that matter, a genomic passenger is limited at best and might be incorrect. Therefore, this must be an iterative learning process in which the results of downstream biological validation and mechanistic studies — and even of clinical experiences from which inhibitors or biomarkers are developed and used — can and must inform the integrative analyses and the validation approaches. This effort will be facilitated by the development or assembly of model systems that are characterized to the same degree as primary tumours and that can be used to quickly test hypotheses suggested by 'omic' analyses of tumours.

For the efficient translation of cancer genome information into the clinic, studies must go beyond statistical analyses of large genomic data sets. This process will require the amalgamation of expertise and insights from cancer biology, cancer genetics, cancer modelling and systems biology, as well as clinical experiences. We suggest that this integrative process will be facilitated by establishing international centres or cooperatives that organize the information obtained from diverse genomic, biological and clinical studies in ways that guide functional analyses and optimize the translation of the cancer genome into effective biomarkers or therapeutics. ■

1. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
2. Collins, F. S. & Barker, A. D. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci. Am.* **296,** 50–57 (2007).
3. Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350,** 2129–2139 (2004).
4. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304,** 1497–1500 (2004).
5. Pao, W. *et al.* EGF receptor gene mutations are common in lung cancers from 'never smokers' and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl Acad. Sci. USA* **101,** 13306–13311 (2004).
   References 3–5 show that a subset of patients with lung cancer have *EGFR* mutations and are responsive to an EGFR-specific tyrosine kinase inhibitor, a finding based on prospective analyses of retrospective data.
6. Druker, B. J. *et al.* Efficacy and safety of a specific inhibitor of the BCR–ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344,** 1031–1037 (2001).
   This paper provides the first proof of concept of targeted therapy: chronic myeloid leukaemia harbouring *BCR–ABL* was shown to be sensitive to treatment with a BCR-ABL-specific tyrosine-kinase inhibitor, imatinib mesylate.
7. Pegram, M. & Slamon, D. Biological rationale for HER2/neu (c-erbB2) as a target for monoclonal antibody therapy. *Semin. Oncol.* **27,** 13–19 (2000).
8. Malkin, D. *et al.* Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250,** 1233–1238 (1990).
9. Futreal, P. A. *et al.* BRCA1 mutations in primary breast and ovarian carcinomas. *Science* **266,** 120–122 (1994).
10. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266,** 66–71 (1994).
11. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378,** 789–792 (1995).
12. Marra, G. & Boland, C. R. Hereditary nonpolyposis colorectal cancer: the syndrome, the genes, and historical perspectives. *J. Natl Cancer Inst.* **87,** 1114–1125 (1995).
13. Gruis, N. A. *et al.* Homozygotes for CDKN2 (p16) germline mutation in Dutch familial melanoma kindreds. *Nature Genet.* **10,** 351–3 (1995).
14. Nowell, P. C. Discovery of the Philadelphia chromosome: a personal perspective. *J. Clin. Invest.* **117,** 2033–2035 (2007).
15. Nardi, V., Azam, M. & Daley, G. Q. Mechanisms and implications of imatinib resistance mutations in BCR–ABL. *Curr. Opin. Hematol.* **11,** 35–43 (2004).
16. Quintas-Cardama, A., Kantarjian, H. & Cortes, J. Flying under the radar: the new wave of BCR-ABL inhibitors. *Nature Rev. Drug Discov.* **6,** 834–848 (2007).
17. Demetri, G. D. Targeting c-kit mutations in solid tumors: scientific rationale and novel therapeutic options. *Semin. Oncol.* **28,** 19–26 (2001).
18. Curtin, J. A., Busam, K., Pinkel, D. & Bastian, B. C. Somatic activation of KIT in distinct subtypes of melanoma. *J. Clin. Oncol.* **24,** 4340–4346 (2006).
19. Hodi, F. *et al.* A major response to Imatinib mesylate in *KIT* mutated melanoma. *J. Clin. Orthod.* (in the press).
20. Rowley, J. D. The role of chromosome translocations in leukemogenesis. *Semin. Hematol.* **36,** 59–72 (1999).
21. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–648 (2005).
22. Bradford, T. J., Tomlins, S. A., Wang, X. & Chinnaiyan, A. M. Molecular markers of prostate cancer. *Urol. Oncol.* **24,** 538–551 (2006).
23. Volik, S. *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA* **100,** 7696–7701 (2003).
24. Bignell, G. R. *et al.* Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **17,** 1296–1303 (2007).
25. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* (in the press).
26. Schechter, A. L. *et al.* The *neu* oncogene: an *erb-B*-related gene encoding a 185,000-M_r tumour antigen. *Nature* **312,** 513–516 (1984).
27. King, C. R., Kraus, M. H. & Aaronson, S. A. Amplification of a novel v-*erbB*-related gene in a human mammary carcinoma. *Science* **229,** 974–976 (1985).
28. Semba, K., Kamata, N., Toyoshima, K. & Yamamoto, T. A v-*erbB*-related protooncogene, c-*erbB*-2, is distinct from the c-*erbB*-1/epidermal growth factor-receptor gene and is amplified in a human salivary gland adenocarcinoma. *Proc. Natl Acad. Sci. USA* **82,** 6497–6501 (1985).
29. Coussens, L. *et al.* Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with *neu* oncogene. *Science* **230,** 1132–1139 (1985).
30. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/*neu* oncogene. *Science* **235,** 177–182 (1987).
   This study correlated *ERBB2* amplification with outcome for individuals with breast cancer.
31. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818–821 (1992).

32. Cameron, D. *et al.* A phase III randomized comparison of lapatinib plus capecitabine versus capecitabine alone in women with advanced breast cancer that has progressed on trastuzumab: updated efficacy and biomarker analyses. *Breast Cancer Res. Treat.* doi:10.1007/s10549-007-9885-0 (in the press).
33. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417,** 949–954 (2002).
34. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304,** 554 (2004).
35. Carpten, J. D. *et al.* A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. *Nature* **448,** 439–444 (2007).
36. Stephens, P. *et al.* Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* **431,** 525–526 (2004).
37. Sharma, S. V., Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nature Rev. Cancer* **7,** 169–181 (2007).
38. Blackhall, F., Ranson, M. & Thatcher, N. Where next for gefitinib in patients with lung cancer? *Lancet Oncol.* **7,** 499–507 (2006).
39. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434,** 917–921 (2005).
40. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
41. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
   References 40 and 41 report on large-scale sequencing studies aimed at identifying somatic mutations in human cancers.
42. Sharpless, N. E. *INK4a/ARF*: a multifunctional tumor suppressor locus. *Mutat. Res.* **576,** 22–38 (2005).
43. Shayesteh, L. *et al. PIK3CA* is implicated as an oncogene in ovarian cancer. *Nature Genet.* **21,** 99–102 (1999).
44. Horvitz, H. R., Shaham, S. & Hengartner, M. O. The genetics of programmed cell death in the nematode *Caenorhabditis elegans. Cold Spring Harb. Symp. Quant. Biol.* **59,** 377–385 (1994).
45. Nurse, P., Masui, Y. & Hartwell, L. Understanding the cell cycle. *Nature Med.* **4,** 1103–1106 (1998).
46. Schreiber-Agus, N. *et al. Drosophila Myc* is oncogenic in mammalian cells and plays a role in the diminutive phenotype. *Proc. Natl Acad. Sci. USA* **94,** 1235–1240 (1997).
47. Kim, M. *et al.* Comparative oncogenomics identifies *NEDD9* as a melanoma metastasis gene. *Cell* **125,** 1269–1281 (2006).
48. Zender, L. *et al.* Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell* **125,** 1253–1267 (2006).
   References 47 and 48 show the power of cross-species integration of cancer genome data for oncogene discovery.
49. Maser, R. S. *et al.* Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* **447,** 966–971 (2007).
   This paper compares the genomes of mouse tumour cells with genetically engineered chromosomal instability to the genomes of various human cancers and shows that there is a significant non-random number of syntenic events, proving that mouse and human cells can experience common biological processes driven by orthologous genetic events during transformation.
50. Sweet-Cordero, A. *et al.* An oncogenic *KRAS2* expression signature identified by cross-species gene-expression analysis. *Nature Genet.* **37,** 48–55 (2005).
51. O'Neil, J. *et al.* Activating *Notch1* mutations in mouse models of T-ALL. *Blood* **107,** 781–785 (2006).
52. Chin, L., Garraway, L. A. & Fisher, D. E. Malignant melanoma: genetics and therapeutics in the genomic era. *Genes Dev.* **20,** 2149–2182 (2006).
53. Hodgson, J. G. *et al.* Copy number aberrations in mouse breast tumors reveal loci and genes important in tumorigenic receptor tyrosine kinase signaling. *Cancer Res.* **65,** 9695–9704 (2005).
54. Artandi, S. E. & DePinho, R. A. Mice without telomerase: what can they teach us about human cancer? *Nature Med.* **6,** 852–855 (2000).
55. Artandi, S. E. *et al.* Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406,** 641–645 (2000).
56. Maser, R. S. *et al.* DNA-dependent protein kinase catalytic subunit is not required for dysfunctional telomere fusion and checkpoint response in the telomerase-deficient mouse. *Mol. Cell. Biol.* **27,** 2253–2265 (2007).
57. O'Hagan, R. C. *et al.* Telomere dysfunction provokes regional amplification and deletion in cancer genomes. *Cancer Cell* **2,** 149–155 (2002).
58. Palomero, T. *et al.* Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. *Nature Med.* **13,** 1203–1210 (2007).
59. Ewart-Toland, A. *et al.* Identification of *Stk6/STK15* as a candidate low-penetrance tumor-susceptibility gene in mouse and human. *Nature Genet.* **34,** 403–412 (2003).
60. Ewart-Toland, A. *et al.* Aurora-*A/STK15 T+91A* is a general low penetrance cancer susceptibility gene: a meta-analysis of multiple cancer types. *Carcinogenesis* **26,** 1368–1373 (2005).
61. Uren, A. G., Kool, J., Berns, A. & van Lohuizen, M. Retroviral insertional mutagenesis: past, present and future. *Oncogene* **24,** 7656–7672 (2005).
62. Sharpless, N. E. & Depinho, R. A. The mighty mouse: genetically engineered mouse models in cancer drug development. *Nature Rev. Drug Discov.* **5,** 741–754 (2006).
63. Westbrook, T. F. *et al.* A genetic screen for candidate tumor suppressors identifies REST. *Cell* **121,** 837–848 (2005).
64. Boehm, J. S. *et al.* Integrative genomic approaches identify *IKBKE* as a breast cancer oncogene. *Cell* **129,** 1065–1079 (2007).
   References 63 and 64 integrate hits from forward genetic screening, using RNAi, with genomic profiles of human cancers to find previously unidentified oncogenes.
65. Berns, K. *et al.* A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumasb resistance in breast cancer. *Cancer Cell* **12,** 395–402 (2007).
66. Staunton, J. E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA* **98,** 10787–10792 (2001).

67. Bild, A. H. *et al*. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–357 (2006).

68. Konecny, G. E. *et al*. Activity of the dual kinase inhibitor lapatinib (GW572016) against HER-2-overexpressing and trastuzumab-treated breast cancer cells. *Cancer Res.* **66,** 1630–1639 (2006).

69. Hieronymus, H. *et al*. Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10,** 321–330 (2006).

70. Wei, G. *et al*. Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* **10,** 331–342 (2006).

71. Neve, R. M. *et al*. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10,** 515–527 (2006).
    This study shows that the cancer genomes of a panel of human cancer cell lines reflect the genomic diversity of human cancers.

72. Wong, K. K. HKI-272 in non small cell lung cancer. *Clin. Cancer Res.* **13,** s4593–s4596 (2007).

73. Scappini, B. *et al*. Changes associated with the development of resistance to imatinib (STI571) in two leukemia cell lines expressing p210 Bcr/Abl protein. *Cancer* **100,** 1459–1471 (2004).

74. Furnari, F. B. *et al*. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev.* **21,** 2683–2710 (2007).

75. Mellinghoff, I. K. *et al*. Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. *N. Engl. J. Med.* **353,** 2012–2024 (2005).

76. Stommel, J. M. *et al*. Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. *Science* **318,** 287–290 (2007).

77. Greshock, J. *et al*. A comparison of DNA copy number profiling platforms. *Cancer Res.* **67,** 10173–10180 (2007).

78. Korbel, J. O. *et al*. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).

79. Drmanac, R. *et al*. DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing. *Science* **260,** 1649–1652 (1993).

80. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).

81. Margulies, M. *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).

82. Shendure, J. *et al*. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).

83. Porreca, G. J. *et al*. Multiplex amplification of large sets of human exons. *Nature Methods* **4,** 931–936 (2007).

84. Costello, J. F. *et al*. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nature Genet.* **24,** 132–138 (2000).

85. Dai, Z. *et al*. An *Asc*I boundary library for the studies of genetic and epigenetic alterations in CpG islands. *Genome Res.* **12,** 1591–1598 (2002).

86. Plass, C. *et al*. An arrayed human *not* I-*Eco*RV boundary library as a tool for RLGS spot analysis. *DNA Res.* **4,** 253–255 (1997).

87. van Steensel, B. & Henikoff, S. Epigenomic profiling using microarrays. *Biotechniques* **35,** 346–350, 352–354, 356–357 (2003).

88. Meissner, A. *et al*. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33,** 5868–5877 (2005).

89. Hu, M. *et al*. Distinct epigenetic changes in the stromal cells of breast cancers. *Nature Genet.* **37,** 899–905 (2005).

90. Leary, R. J., Cummins, J., Wang, T. L. & Velculescu, V. E. Digital karyotyping. *Nature Protoc.* **2,** 1973–1986 (2007).

91. Collas, P. & Dahl, J. A. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front. Biosci.* **13,** 929–943 (2008).

Enabling personalized cancer medicine through analysis of gene-expression patterns

Laura J. van 't Veer[1,2,3] & René Bernards[1,3,4,5]

Therapies for patients with cancer have changed gradually over the past decade, moving away from the administration of broadly acting cytotoxic drugs towards the use of more-specific therapies that are targeted to each tumour. To facilitate this shift, tests need to be developed to identify those individuals who require therapy and those who are most likely to benefit from certain therapies. In particular, tests that predict the clinical outcome for patients on the basis of the genes expressed by their tumours are likely to increasingly affect patient management, heralding a new era of personalized medicine.

The conventional approach to cancer therapy has been to provide treatment according to the organ or tissue in which the cancer originates. This approach was appropriate when there was only a rudimentary understanding of the molecular origins of cancer and the different intracellular signalling pathways that are perturbed in the various types of cancer (such as in breast cancer or lung cancer). In the past two to three decades, however, the genetic events that lead to cancer have been dissected, and it has become clear that cancer develops as a result of multiple genetic defects and that individuals with the same type of cancer often have dissimilar genetic defects in their tumours. This finding explains why patients who seem to have similar cancers respond in a heterogeneous manner to anticancer agents and shows clearly the huge obstacle to providing effective treatments for cancer.

In the past decade, cancer therapy has slowly but steadily begun to shift from a 'one size fits all' approach to a more personalized approach, in which each patient is treated according to the specific genetic defects in the tumour. Such an individualized approach requires the discovery and development of biomarkers (biological indicators) that help doctors to decide which patients to treat (known as prognostic biomarkers) and which therapy is most likely to be effective for a given patient (known as predictive biomarkers). More specifically, prognostic biomarkers predict the clinical outcome for a patient if no anticancer drugs are administered, whereas predictive biomarkers predict the outcome of a specific therapy for a patient. An example of why such biomarkers are needed to improve patient management is that, for some tumours, resection (that is, surgical removal) of the primary tumour might be curative; therefore, systemic therapy to eliminate any remaining tumour cells (also known as adjuvant therapy) would not be needed. By contrast, for more malignant primary tumours, aggressive systemic therapy, often chemotherapy, might be required after resection, in order to reduce the risk of the tumour recurring. However, the distinction between these is often unclear, so prognostic biomarkers that enable the likelihood of recurrence to be determined are urgently needed in the clinic.

In the case of breast cancer, large meta-analyses have shown that recurrence is likely in 20–30% of young women with early-stage (lymph-node-negative) breast cancer who undergo only surgery and localized radiation treatment[1]. But, in the United States, 85–95% of women with this type of cancer receive adjuvant chemotherapy, mostly because conventional clinicopathological parameters fail to identify reliably those patients who

are likely to relapse. Therefore, 55–75% of women with early-stage breast cancer in the United States undergo a toxic therapy from which they will not benefit but will experience the side effects. So it is not surprising that the initial attempts to discover clinically relevant prognostic biomarkers have focused on breast cancer (discussed later).

The advent of DNA-microarray technology in the 1990s (refs 2, 3) made it possible to assess the expression of tens of thousands of genes in a single experiment. Systematic analysis of the gene-expression patterns of tumour samples enabled researchers to identify characteristic expression patterns of groups of genes that are associated with specific tumour traits. These patterns are known as gene-expression signatures.

In this review, we focus on gene-expression signatures as a new class of molecular diagnostic test for cancer. We discuss pitfalls in the discovery of gene-expression signatures, how such signatures can be used to develop clinically relevant tests and how these tests are likely to affect patient management and drug development in the future.

Building gene-expression profiles

The massive parallel quantification of messenger RNA abundance that is possible using DNA-microarray technology has enabled genome-wide gene-expression data to be collected for large numbers of biological specimens. Collecting this unprecedented amount of data (at least for biologists) has necessitated the development of new tools to analyse the large data sets. In principle, to find connections between the patterns of gene expression by tumour cells and the behaviour of these cells, there are three approaches: the data-driven approach, the knowledge-driven approach and the model-driven approach.

The most straightforward is the data-driven approach, in which a genome-wide analysis of gene expression is carried out, and then correlates between patterns of gene expression and certain tumour traits are searched for. The strength of this approach is that it is unbiased: there are no assumptions about which genes are likely to be involved in the process of interest. For example, in a data-driven study of the prognosis of patients with breast cancer, little was known about the function of 15 of the 70 genes that were found to constitute a prognostic gene-expression signature[4]. A drawback of this approach is that the outcome relies solely on the quality of the data (and the samples).

By contrast, using the knowledge-driven approach, genes that are thought to be relevant to a particular cancer trait are selected on the basis

of the scientific literature. This approach is often used when only forma-lin-fixed paraffin-embedded tumour tissue is available. The RNA isolated from such tissue is fragmented, and such poor-quality RNA is far from ideal for genome-wide quantitative analysis using DNA microarrays. It can, however, be analysed by PCR with reverse transcription. In this approach, mRNA abundance is measured by using unique pairs of oligonucleotide primers that correspond to each gene, a labour-intensive process that precludes genome-scale analysis of gene expression. Hence, in studies involving formalin-fixed paraffin-embedded material, sets of 'likely suspect' genes are often tested. A drawback of this approach is that the outcome is only as good as the state of knowledge: genes that are not known to be involved in a process cannot be considered.

In the model-driven approach, the transcriptional responses of cells after exposure to specific stimuli are used to predict tumour traits. For example, a gene-expression signature for wound healing has been used to predict the survival of individuals with breast cancer[5]. Similarly, gene-expression signatures that reflect the activation of specific oncogenic pathways have been used both to determine prognosis and to predict responses to anticancer drugs[6,7]. This approach has the drawback that the experimental model used might not accurately reflect the processes that occur in tumours.

A comparison of the data-driven approach and the knowledge-driven approach applied to a similar problem sheds more light on the advantages and disadvantages of these approaches. If the presence of a certain transcription factor is known to affect the prognosis of individuals with a particular cancer, then in the knowledge-driven approach to building a prognostic gene-expression signature, the gene encoding this transcription factor would be incorporated into a prognostic signature. In some cancers, however, this gene might be expressed, but its product non-functional (for example, as a result of a missense mutation). For this reason, in a data-driven approach, targets downstream of a transcription factor of interest are often found to be distinguishing features, rather than the gene encoding the transcription factor itself, because the expression of these targets provides more relevant information on the activity of the transcription factor. As an example, a 16-gene signature for the prognosis of breast cancer that was derived from 250 'candidate' genes selected on the basis of published studies includes *ESR1,* which encodes oestrogen receptor-α (ER-α; a transcription factor that is expressed by most breast cancers)[8]. By contrast, a set of 70 genes for assessing breast-cancer prognosis (discussed earlier) that was identified by a data-driven approach does not include *ESR1* itself but includes several genes that are targets of ER-α[4].

After large-scale gene-expression data sets have been collected, there are two fundamentally different ways to analyse them. One approach is to ask whether in a group of samples (for example, tissues from individuals diagnosed with a particular cancer), there are subgroups (or clusters) of samples with similar gene-expression patterns. These similarities in gene expression can be used to classify a cancer into subtypes that could have similarities in biological behaviour. This type of data analysis is called hierarchical clustering or unsupervised classification[9], and it has the advantage that additional clinical data are not required. For example, hierarchical clustering of breast-cancer specimens[10,11] identified five naturally occurring subtypes (referred to as 'intrinsic subtypes') of breast cancer that had not been observed previously. Some of these intrinsic subtypes differ markedly in their aggressiveness; for example, the prognosis for individuals with the subtype called basal-like breast cancer is significantly worse than that for individuals with another subtype, luminal-type breast cancer[11,12].

The second approach to data analysis is known as supervised classification. Samples are divided into groups that are known to have different clinical end points (for example, recurrence versus no recurrence, and drug response versus no drug response), and genes that can correctly identify the distinct groups are searched for (Fig. 1). One set of tumours (called the training set) is used to identify the genes that discriminate between the groups — the gene-expression signature — and then a second, independent, set of tumours (called the validation set) is used to test how well these genes can classify samples that have not been

grouped. The situations in which these approaches have been used and their relative value are discussed in the next section.

Prognostic profiles

Initial studies of gene-expression profiling of cancer used simple hierarchical clustering to identify subtypes among apparently similar cancers. A landmark early study identified two distinct subtypes of diffuse large B-cell lymphoma (DLBCL) — germinal-centre B-like DLBCL, and activated B-like DLBCL — for which the overall survival of patients significantly differs[13]. This was the first in an avalanche of publications showing that the molecular classification of tumours on the basis of
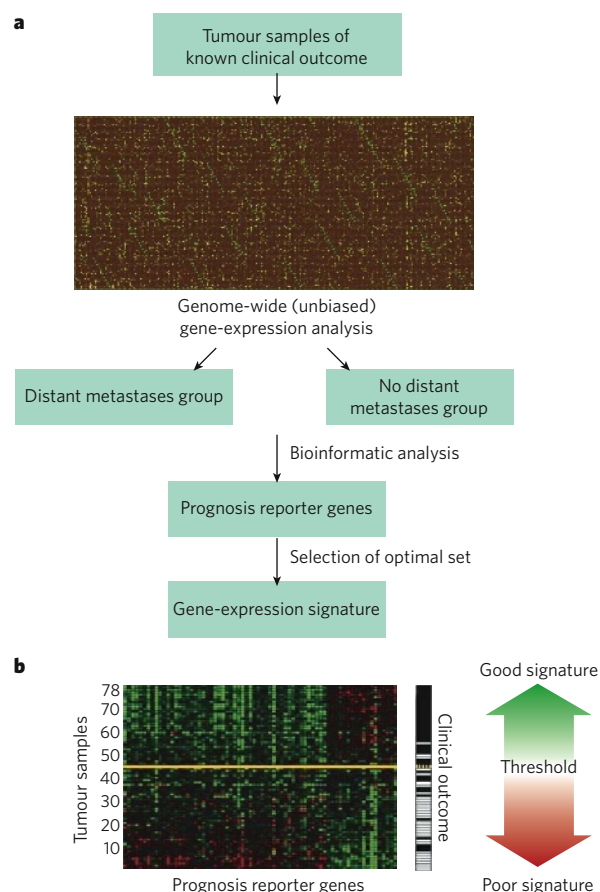

Figure 1 | Predicting disease outcome by using complex gene-expression tests. a, Generating a prognostic gene-expression signature by using supervised classification. The gene expression of cells in a set of tumours of known clinical outcome is analysed by using whole-genome microarrays. Colours indicate the level of expression for each gene: red, gene is more active than the average for tumours of this type; green, gene is less active than average; yellow, gene is equally active; and, black, gene is not expressed. The results for each tumour sample are then classified into two categories: tumours with a good outcome (no distant metastases developed), and tumours with a poor outcome (distant metastases developed). Using bioinformatic analysis, genes whose expression is significantly correlated with disease outcome are identified, and these are known as prognosis reporter genes. An optimal set of genes is then selected from the prognosis reporter genes by using bioinformatic algorithms, and the pattern of expression of this multigene set is known as a gene-expression signature (or classifier). **b,** The gene-expression signature generated in **a** is shown as a 'heat map'. The expression of the 70 prognosis reporter genes selected as the optimal set (vertical columns) is shown for 78 tumours (horizontal lines). So each of the 70×78 intersection points of the heat map shows how a particular gene is expressed in a given tumour. A red spot indicates that the gene is expressed at a higher level than average for tumours of this type, and a green spot that the gene is expressed at a lower level. The outcome of the disease is shown on the right: white indicates metastasis; black indicates no metastasis; and yellow indicates the threshold for metastasis. (Panel adapted, with permission, from ref. 4.)
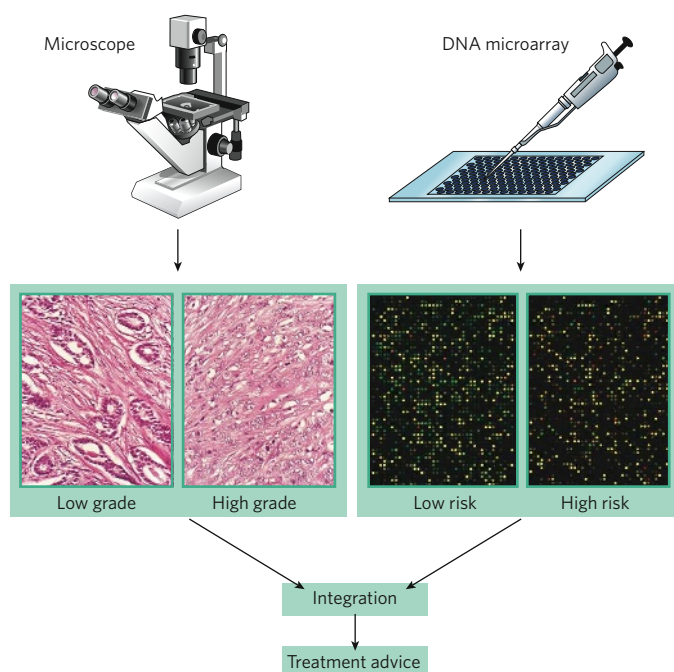
565

Figure 2 | Conventional and molecular diagnostic testing for cancer. Conventional diagnostic tests rely heavily on morphological criteria (that is, properties of cells that can be observed microscopically in tissue sections from a tumour) to judge the aggressiveness of cancer, a process known as grading. As an example, the differences between a 'low grade' adenocarcinoma of the breast (for which patients generally have a favourable prognosis) and a 'high grade' adenocarcinoma of the breast (for which patients have a worse prognosis) can be observed from the images. More recently, multigene-expression tests (also known as *in vitro* diagnostic multivariate index assays, IVDMIAs) have been shown to be powerful tools for predicting disease outcome and have become subject to scrutiny by the US Food and Drug Administration. As an example, a molecular description of low-risk and high-risk adenocarcinoma of the breast, as judged by a gene-expression signature consisting of 70 genes, is shown. One current challenge is how to integrate the knowledge obtained from these conventional tests and molecular diagnostic tests into a single recommendation for the oncologist treating the patient.

gene expression can identify previously undetected and clinically significant subtypes of cancer. Even though the gene-expression signatures found in these studies uncovered aspects of tumour-cell biology that had gone unnoticed, these studies were not designed to find signatures identifying subtypes of cancer that result in different prognoses. Their main purpose was to establish a molecular classification of tumours on the basis of their gene-expression patterns.

In a study that was designed to find a prognostic gene-expression signature for breast cancer, gene-expression data from breast-tumour samples of known clinical outcome were analysed by using supervised classification[4]. This study yielded the previously mentioned 70-gene signature for breast-cancer prognosis. Remarkably, when another research group independently used a similar approach[14], a 76-gene signature was identified, but this had only three genes in common with the 70-gene signature. This finding was interpreted by some to indicate that such gene-expression signatures are highly unstable[15]. But, because both of these gene-expression signatures could be independently validated in large groups of patients[16,17], it is more likely that the two signatures use different genes to monitor the same biological processes, an idea for which there is supporting evidence[18,19]. Using similar strategies, gene-expression signatures that assess the risk of recurrence of non-small-cell lung cancer[20] and several other cancer types have also been established.

A combination of a knowledge-based approach and a data-driven approach was used to identify the previously mentioned 16-gene signature for breast-cancer prognosis. First, a set of 250 candidate genes

was identified and, after analysing the expression of the candidate genes in about 400 tumours, 16 genes were selected from this set[8]. Of all the gene-expression signatures for cancer that have been identified, only three are commercially available: the 70-gene signature for breast-cancer prognosis is available under the name MammaPrint (Agendia); the 16-gene signature as Oncotype DX (Genomic Health); and a 2-gene signature[21], which has recently been released, under the name the H/I test (AviaraDx). Now that such tests can be used in the clinic, one of the major challenges facing oncologists and pathologists is how to integrate the information obtained from conventional tests with that from these molecular tests (Fig. 2).

An important question when considering the development of new tests is whether prognostic gene-expression profiles are independent of the molecularly defined subtypes of cancer. As mentioned earlier, individuals with basal-like breast cancers generally have a worse prognosis than those with luminal-type breast cancers. Should a separate prognostic signature therefore be made for the basal-like subtype and the luminal-type subtype, or is a signature based on supervised classification of a diverse panel of breast tumours equally powerful for assessing the prognosis of patients with either subtype? Our studies of tumours from 295 patients with breast cancer indicate that within the group of patients who were predicted to be at high risk of relapse by using the 70-gene signature, the patients with the basal-like subtype have a comparable outcome to those with the luminal-type subtype (L.J.v.V., unpublished observations). This finding suggests that the prognostic value of the molecular subtyping of cancer, as was carried out in the initial gene-expression profiling studies, has been surpassed by that of prognostic gene-expression signatures such as the 70-gene signature and that the underlying molecular subtype does not contain additional crucial information for determining a patient's prognosis. It is possible, however, that patients with the various molecular subtypes of cancer respond differently to particular therapies.

Predictive profiles

In the area of predicting responses to particular therapies, gene-expression profiling studies have not yet delivered on their promise. It seems that responses to anticancer drugs are more difficult to predict by using molecular tests than prognosis is. One of the main reasons for this difficulty is that resistance to anticancer agents can result from a variety of mechanisms. Consequently, there might not be a gene-expression profile that correlates with resistance to a certain drug. This is true for both resistance to conventional chemotherapeutic agents, which are often pleiotropic, and to newer 'targeted therapies', which affect specific components of signalling pathways. It should be pointed out, however, that the differences between these types of anticancer drug are less marked than is generally assumed: some of the conventional chemotherapeutic agents (such as topoisomerase inhibitors) target specific enzymes, whereas some of the small-molecule-based targeted therapies (such as imatinib mesylate and lapatinib) are directed against more than one enzyme. In addition, resistance to drugs might result from subtle mutations that do not cause gross changes in gene expression, a process that is therefore undetectable by gene-expression profiling.

Another important impediment to the discovery of predictive gene-expression signatures is that, for genome-wide gene-expression studies, large numbers of tumour samples are required (of the order of 100), to reduce the probability that associations between gene-expression signatures and therapy outcomes are spurious. An additional consideration is that drug-response gene-expression profiles cannot be constructed from tumour samples from patients who have undergone adjuvant therapy. Patients who do not relapse after adjuvant therapy could have had a tumour with either a 'good outcome' (that is, a tumour that did not metastasize, for which adjuvant therapy was unnecessary) or a 'poor outcome' (that is, a tumour that metastasized and, in this case, responded to therapy). Hence, prediction cannot be separated from prognosis in this type of study. In principle, tumour samples from patients who are known to have metastatic disease could be used to develop drug-response profiles (all are, by definition, of the poor-outcome type), but such samples are not readily available, because patients who are treated for metastatic cancer

do not usually have their tumours biopsied first. Furthermore, patients with metastatic cancer often receive combination therapies, making it difficult to determine which drug a patient responded to. One way to circumvent some of these issues is to use gene-expression profiling in a 'neo-adjuvant' setting, in which patients are treated systemically before the resection of a primary tumour (often carried out when the tumour is large). Short-term responses to anticancer drugs can be determined in this setting by using imaging technologies, and primary tumours can often be sampled by needle biopsy[22,23].

As a result of all of these factors, few drug-response gene-expression signatures have been published, and those that have are not clinically useful in their current form because they have not been properly validated against an independent series of tumours. It should be noted, however, that many clinical trials exploring the neo-adjuvant approach are underway, a well-documented example being the I-SPY trial, sponsored by the National Cancer Institute (http://tr.nci.nih.gov/iSpy).

A more general problem with predictive gene-expression signatures is that doctors often do not have a suitable replacement for the first-line therapy and are unlikely to withhold treatment entirely just because a biomarker indicates that the treatment probably will not be effective. Thus, the sensitivity (that is, the percentage of patients with a certain trait that test positive) and the specificity (that is, the percentage of patients without a certain trait that test negative) of predictive gene-expression signatures will need to improve markedly before they will be useful in the clinic.

Potential short cuts to predictive biomarker development

Given the problems with generating predictive gene-expression profiles, several groups have attempted to generate such profiles through identifying genes or pathways that potentially affect how a cell responds to a drug, often by using models based on cell lines. A small number of human tumour samples can then be tested for the expression of these *in vitro*-generated sets of candidate genes. The approaches that have been explored so far are discussed in this section and illustrated in Fig. 3.

In vitro-generated predictive profiles using cell-line models

Gene-expression profiling of a series of human cancer cell lines with known drug sensitivity has been used to identify patterns of gene expression that correlate with responses to drugs *in vitro*[24–26]. These studies used a heterogeneous set of cell lines (a panel of 60 human cancer cell lines of different tissue origins, called NCI-60) with known sensitivity profiles to several anticancer drugs. The gene-expression profiles generated are therefore designed to be independent of the tissue of origin, but it is questionable whether such tissue-independent profiles are reliable predictors of drug responses. The validity of this approach has been challenged recently[27]; however, an independent validation of such *in vitro*-generated gene-expression signatures was also published around the same time[28]. It might be more fruitful to carry out such gene-expression profiling studies on panels of cell lines derived from a single type of cancer. Using a more homogeneous panel of cell lines might allow the identification of drug-resistance mechanisms that are missed when a heterogeneous set of cancer cell lines is used. Indeed, for breast cancer, the first results obtained by using this approach seem to be promising[29]. But there is no direct evidence that using a more homogeneous cell-line panel yields a better profile. An example of the general approach is shown in Fig. 3a.

Signalling pathways

In the past, the most effective therapies for cancer were based on empirically derived evidence from large clinical studies. This approach has resulted in a standard therapeutic regimen in which drug combination A is given for breast cancer, for example, but combination B is given for lung cancer, and so on. The continued use of such standard protocols to treat each type of cancer ignores one of the most important lessons from the past two decades of molecular genetics research on cancer: namely, that each tumour has a complex and unique set of genetic alterations that drive the oncogenic proliferation of the cells. Nevertheless, the fact that distinct drug combinations show activity against specific types of cancer reflects differences in the molecular pathways that are predominantly
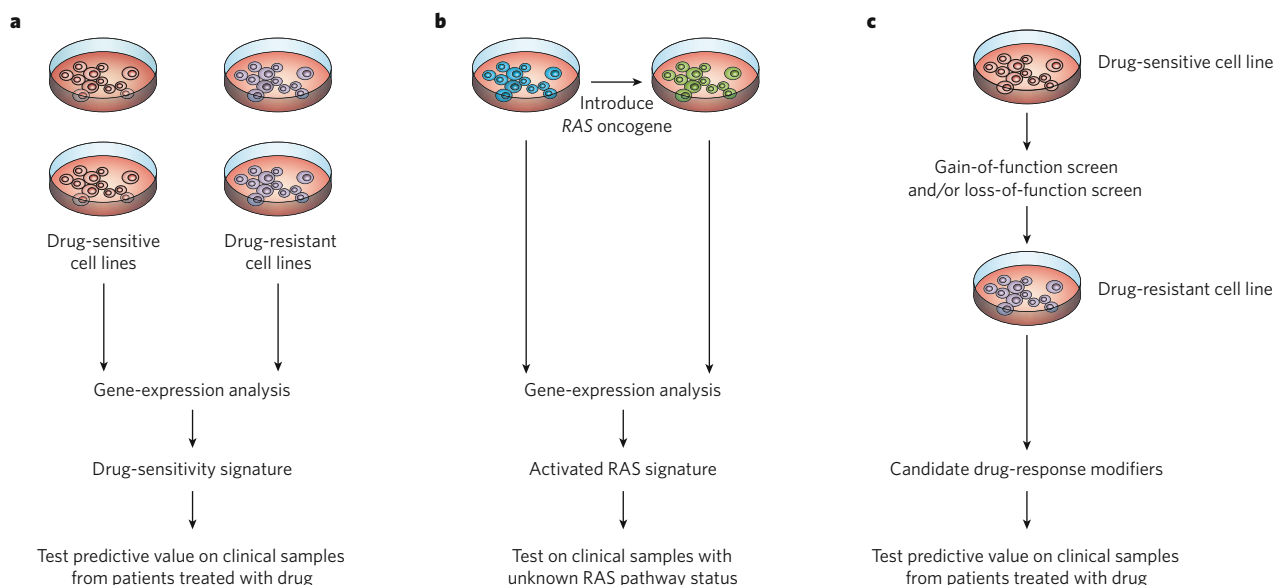

Figure 3 | Short cuts to the development of drug-response biomarkers.
a, Collections of tumour cell lines of known drug sensitivity can be used to build gene-expression signatures that discriminate between sensitive and resistant cell lines. Such *in vitro*-generated drug-sensitivity signatures can be validated on tumour samples from patients treated with the same drugs. **b**, Gene-expression signatures for signalling pathways can be constructed *in vitro* by introducing the gene of interest (a mutant *RAS* gene that is constitutively active in the example here) into tumour cell lines and studying the effect of the presence of the oncogene on genome-wide gene expression. Tumour samples for which the status of the RAS pathway is unknown can then be assessed by comparing their gene-expression patterns with that of the 'activated RAS pathway' identified *in vitro*. If a drug that

targets the RAS pathway is available, then similarity between the gene-expression profile of the tumour and a RAS pathway signature could be used to guide the choice of therapy. **c**, Functional genetic approaches can be used *in vitro* to uncover which genes can contribute to drug resistance in tumour cell lines. More specifically, using these approaches — genome-scale gain-of-function screens or RNA-interference-based loss-of-function screens — full-length complementary DNAs or small interfering RNAs are introduced to change the abundance of gene products, turning drug-sensitive cell lines into drug-resistant cell lines. The predictive ability of the genes that are candidates for modifying drug responses can then be examined by assessing their expression levels in a relatively small number of clinical samples from patients treated with the same drug.

deregulated in these cancers. As such, these empirically derived treatment protocols can be viewed as primitive forms of targeted therapy, supporting the idea of stratifying tumours mainly according to their signalling-pathway perturbations rather than their tissue of origin. The challenge in developing molecular tests to predict drug responses is to identify the altered pathways in each tumour so that each patient receives the optimal targeted therapy. To facilitate this, there needs to be a shift away from describing cancers according to their tissue and cell type of origin (for example, adenocarcinoma of the breast) towards describing them by the main pathways that drive tumour-cell proliferation (for example, phosphatidylinositol-3-OH kinase (PI(3)K)-driven cancer or WNT-driven cancer).

Indeed, for more than 30 years, it is has been routine practice to classify breast tumours as ER positive or ER negative, and this classification is commonly used to decide a patient's eligibility for hormonal therapy, which is one of the main forms of adjuvant therapy for breast cancer (involving inhibition of the growth-stimulating effects of the female hormone oestrogen on the cancer). Therefore, neither the concept of targeted therapy nor the concept of naming tumours by the pathways that drive their proliferation is new. The main reason why it has not been adopted on a larger scale is because the tools to measure the activation of various signalling pathways have been lacking. Moreover, for most cancers, it is unclear which pathway drives the oncogenic process.

One important recent advance that should aid the development of tests to predict drug responses is the finding that activation of a signalling pathway leads to characteristic changes in gene expression, which can be identified using gene-expression analysis. Joseph Nevins's research group[6,7,30], for example, has established several 'pathway gene-expression signatures' (the RAS signature, SRC signature, MYC signature and E2F signature) by experimentally manipulating cell lines to activate certain pathways, and these signatures can be used both to determine prognosis and to select specific therapies that target the activated pathways. This approach is illustrated in Fig. 3b. Whether such pathway signatures are similar in different tissue types remains to be investigated. It might be necessary to determine pathway signatures for each cancer type so that the optimal treatment can be chosen for each patient.

Such cell-line models are useful for building pathway gene-expression signatures, but they are often not an accurate representation of the *in vivo* situation in human cancer. To overcome this problem, Lao Saal *et al.*[31] analysed human breast-cancer samples by using immunohistochemistry to identify a set of primary tumours that expressed the tumour-suppressor gene *PTEN* (phosphatase and tensin homologue, which is a component of the PI(3)K-signalling pathway) at a low level and then linked this 'PTEN-low' phenotype to a specific gene-expression signature of these tumours. Searching for this signature was shown to be a more sensitive way to detect tumours in which the PTEN–PI(3)K pathway is activated than using immunohistochemistry to detect PTEN itself (because events other than loss of, or a reduction in, *PTEN* expression can also modulate the pathway). Such a 'pathway-integrative' signature might therefore be an effective tool to guide the use of therapies that target this pathway[31].

Alternatively, instead of starting from a cancer specimen, the process can be inverted: first, a database of gene-expression patterns that result from well-defined perturbations of specific pathways can be established, and then a test gene-expression data set (for example, from a given cancer specimen) can be assessed for how well it matches any of the gene-expression profiles in the database[32,33]. Such gene-expression compendia (also known as connectivity maps) can be powerful tools for deciding how to treat patients. This approach uncovered a similarity between the effects of exposure to heat-shock protein 90 inhibitors and inhibition of androgen-receptor signalling, a signalling pathway that is central to prostate cancer[34]. Similarly (and perhaps of greater clinical relevance), searching a database of gene-expression signatures induced in response to drug treatment revealed that rapamycin (which inhibits mTOR) induces a signature that overlaps with a signature for sensitivity to glucocorticoids that is found in samples from patients with acute lymphoblastic leukaemia (ALL)[35]. Indeed, when tested, rapamycin was able to induce glucocorticoid sensitivity in ALL samples. These data

underscore the power of this type of approach to identify connections between apparently unrelated biological perturbations, which can lead to important insights into the factors that mediate drug sensitivity.

Functional genetic approaches
Another way to identify genes that can serve as biomarkers to predict drug responses is to take suitable cell-line models and use genetic screens to find genes or pathways whose altered activity modulates sensitivity to anticancer drugs *in vitro* (Fig. 3c). After such candidate genes have been identified, their expression can be measured in tumour samples from patients with cancer who have been treated with the same drug, and the pattern is then correlated with resistance to the drug. The validation of a defined set of candidate genes requires fewer tumour samples than an unbiased (genome-wide) search for predictive biomarkers, thereby bypassing one of the largest bottlenecks in the discovery of robust biomarkers: the availability of suitable tumour samples.

This approach might be useful to improve responses to trastuzumab (Herceptin), for example. This antibody-based drug is effective, particularly in combination with chemotherapy, for treating patients with breast cancers that produce the epidermal growth-factor receptor ERBB2 (also known as HER2 and NEU). But less than 35% of patients with *ERBB2*-expressing metastatic breast cancer respond when treated with trastuzumab alone, and it is largely unclear why this is the case[36]. Moreover, gene-expression studies of samples from patient who were treated in the neo-adjuvant setting with trastuzumab have failed to find a clear drug-response profile[22]. To uncover trastuzumab-response modifiers, we have used loss-of-function RNA-interference screens, in which gene activity is suppressed on a large scale by using short duplex RNAs, to identify genes that confer resistance to trastuzumab in *ERBB2*-expressing breast-cancer cell lines. Of a set of 8,000 genes tested, only loss of expression of the tumour-suppressor gene *PTEN* conferred resistance to trastuzumab[37], a finding in close agreement with an earlier study implicating this gene in resistance to trastuzumab[38]. Gene-expression signatures for the loss of *PTEN* expression in breast cancer are available[31], and these could be suitable for predicting responses to trastuzumab, as well as to other therapies (such as gefitinib and lapatinib) targeted to ERBB2 or the various members of the epidermal growth-factor receptor family.

Using a similar functional genetic approach, Charles Swanton *et al.*[39] identified regulators of mitotic arrest and ceramide metabolism as determinants of sensitivity to paclitaxel and other chemotherapeutic drugs, and Angelique Whitehurst *et al.*[40] identified genes whose suppression increases the efficacy of paclitaxel against non-small-cell lung cancer. Such synthetic lethal interactions (that is, a combination of two non-lethal events that together result in cell death) could be used to suggest valuable combination therapies for cancer. Along similar lines, two groups recently found that breast tumours with a mutation in *BRCA1* or *BRCA2* are hypersensitive to inhibitors of poly(ADP-ribose) polymerase 1 (PARP1)[41,42]. Gene-expression signatures that identify tumours with defects in the BRCA1 pathway are available[4], so these could be used to determine which patients will respond to PARP1 inhibitors.

Implementing gene-expression profiles in the clinic
Translating biomarker research into clinically useful tests has often been a frustrating activity. Many of the biomarkers identified in the initial tumour studies, which were retrospective, failed to be validated in subsequent studies. One of the main reasons for these failures was that early biomarker discovery was knowledge-driven, but the knowledge was often of poor quality. By contrast, some of the more recent gene-expression signatures were derived from large data-driven, genome-wide studies with excellent data quality, so these biomarkers are far more likely to be validated than previously identified biomarkers from knowledge-driven studies.

Both the US Food and Drug Administration (FDA) and the medical community have recognized that multigene signatures are better biomarkers than single molecules, so why are so few gene-expression signatures available in the clinic? First, on the basis of past failures, doctors are often reluctant to use biomarkers that have been validated only

by retrospective studies; they insist on validation by prospective studies before biomarkers are used in routine clinical practice. A second impediment is that DNA-microarray technology was initially not very robust (at least around the year 2000) and, to many scientists and doctors, it still has a poor reputation, which has been unfounded since industry became involved in production[43]. Third, the correct regulatory path for using multigene tests in clinical practice is unclear. For two of the three multigene tests that are commercially available, large prospective validation studies are in progress: the study TAILORx (ref. 44) is validating the 16-gene signature marketed as Oncotype DX (ref. 8); and the study MINDACT (ref. 45) is validating the 70-gene signature marketed as MammaPrint[4,46]. But the results of these large studies, which require thousands of patients, are at least five years away, and the medical community seems to be divided about when to start using such tests in routine clinical practice[47]. This dilemma was exacerbated recently when the FDA cleared the first multigene assay (MammaPrint) on the basis of only retrospective validation (http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/PMNSimpleSearch.cfm?db=PMN&id=K062694). It should also be considered that the cost of large prospective trials is prohibitive for developing the 'average' molecular diagnostic test. Last, it is often forgotten that purely prognostic gene-expression profiles based on archival tumour samples from patients who were not given adjuvant therapy (which, at least for early-stage breast cancer, is a relatively recent inclusion in the treatment regimen) cannot be validated in current prospective clinical trials, because most patients now receive some form of adjuvant therapy. And, as mentioned earlier, adjuvant therapy 'contaminates' the validation of a purely prognostic profile, so clinical studies in which adjuvant therapy is used are suboptimal for the validation of prognostic gene-expression signatures. For all of these reasons, we expect that robust retrospective studies will become the norm for validating prognostic gene-expression signatures.

The FDA's involvement in regulating molecular diagnostic tests is contested by some. But it seems logical for several reasons that the FDA take an active role in the market approval of these tests. First, molecular diagnostic tests are likely to increasingly affect patient management as tests begin to show that treatment with a particular drug will benefit groups of patients whose tumours have a certain gene-expression pattern. However, if the predictions of such tests are incorrect, then patients could be given an inappropriate drug. Second, these molecular tests are complex, and their accuracy and reproducibility are unlikely to be understood by most doctors. It is therefore improbable that the 'average' doctor would know whether a gene-expression test reliably predicts responses to an anticancer drug or reliably assigns a tumour as having a high risk of recurrence. For these reasons, the FDA has expressed its intention to regulate these molecular diagnostic tests similarly to how it regulates anticancer drugs.

In September 2006, the FDA issued a draft guidance document for the use of a new type of molecular diagnostic test called an *in vitro* diagnostic multivariate index assay (IVDMIA). These tests use complex mathematical algorithms to interpret large amounts of gene- or protein-expression data for the purpose of guiding medical decision-making. After a period open for comment by the diagnostic industry and clinical laboratories, the FDA issued a second version of this draft document in July 2007 (http://www.fda.gov/cdrh/oivd/guidance/1610.pdf). It is expected that the final version of this document will become the basis for regulatory enforcement of the molecular diagnostic industry in the United States. Unfortunately, Europe (once again) lags behind. The European counterpart of the FDA, the European Medicines Agency, has not announced any plans to regulate complex molecular diagnostic tests other than requiring the Conformité Européene (CE) marking (a generic mark that is mandatory for many products on the market in the European Union, indicating that manufacturers have conformed to EU legislation). The regulation of molecular diagnostic tests is viewed by some as being a burden and impeding innovation. Others, however, think that the FDA's approval of molecular diagnostic tests will improve the acceptance of these tests in the clinic. Acceptance might not be far off, however, considering that one in eight women with early-stage breast

cancer in the United States is likely to have a molecular diagnostic gene-expression test in 2008 (based on sales of Oncotype DX). It seems that not all doctors are going to wait for the results of the prospective validation studies that are underway.

Future perspectives

The intention of the FDA to regulate molecular diagnostic tests reflects the increased effect that these tests are likely to have on patient management. Traditionally, diagnostic tests for cancer have been carried out in local hospitals. The quality control for these tests has been poor though. A large study that investigated interlaboratory variance in the immunohistochemical detection of ER in breast cancer, across 200 laboratories in 26 countries, showed a false-negative rate of 30–60% (ref. 48), yet this type of test is used routinely to decide whether a patient should receive hormonal therapy. By contrast, a recent study investigating interlaboratory variance of a DNA-microarray-based test for breast-cancer prognosis found an extremely high concordance between laboratories[49], which is consistent with other studies of the reproducibility of using microarray platforms[43]. Given their prognostic power and reliability, molecular diagnostic tests are expected to become increasingly relevant tools in tailoring care to each patient. Diagnostic tests have traditionally been thought of as low-cost items in the health-care chain; however, the new generation of molecular diagnostic tests will be more drug-like in terms of their effects on patient care, their oversight by regulatory authorities and, consequently, their cost. The increased expenditure on molecular diagnostics could, however, be mitigated by the subsequent reduction in use of costly (molecularly targeted) therapies[50].

Molecular diagnostic tests will also be pivotal in identifying patients who respond to experimental anticancer drugs in clinical trials. Increasingly, drugs will be developed together with a dedicated companion diagnostic test that identifies responders to the drug in question. This co-development of drug and biomarker underscores the need for regulatory authorities to control both the drug and the companion diagnostic test. There seems to be no return from this new path to drug development. At first glance, stratifying patients in this way might seem unattractive to the pharmaceutical industry, because it reduces the size of the market for each anticancer drug. Conversely, however, molecular profiling might uncover commonalities between seemingly different tumours, potentially expanding the market for a candidate drug. When a drug has been developed and is available together with a companion diagnostic test that correctly identifies patients who benefit from the drug, there is no longer a place for a similar drug for which patients cannot be adequately stratified. Or, as one of our colleagues in the drug development industry said recently: "Pharmacogenomics, you either do it, or it is done to you." For once, the choice seems simple.   ∎

1. Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet* **352,** 930–942 (1998).
2. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
3. Fodor, S. P., Read, J. L., Pirrung, M. C., Lu, A. T. & Solas, D. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251,** 767–773 (1991).
4. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).
5. Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA* **102,** 3738–3743 (2005).
6. Bild, A. H., Potti, A. & Nevins, J. R. Linking oncogenic pathways with therapeutic opportunities. *Nature Rev. Cancer* **6,** 735–741 (2006).
7. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–357 (2006).
    This paper shows that the activation status of oncogenic pathways is associated with a characteristic gene-expression signature and that these signatures can be used both to determine a patient's prognosis and to predict his or her response to particular drugs.
8. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351,** 2817–2826 (2004).
9. Quackenbush, J. Computational analysis of microarray data. *Nature Rev. Genet.* **2,** 418–427 (2001).
10. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
11. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98,** 10869–10874 (2001).
12. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100,** 8418–8423 (2003).

13. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511 (2000).
    This study shows that previously undetected and clinically significant subtypes of cancer can be identified by molecular classification of tumours on the basis of gene expression.
14. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365,** 671–679 (2005).
15. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl Acad. Sci. USA* **103,** 5923–5928 (2006).
16. Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* **13,** 3207–3214 (2007).
17. Buyse, M. *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl Cancer Inst.* **98,** 1183–1192 (2006).
18. Fan, C. *et al.* Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.* **355,** 560–569 (2006).
19. Yu, J. X. *et al.* Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer* **7,** 182, doi:10.1186/1471-2407-7-182 (2007).
20. Potti, A. *et al.* A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N. Engl. J. Med.* **355,** 570–580 (2006).
21. Ma, X.-J. *et al.* The *HOXB13:IL17BR* expression index is a prognostic factor in early-stage breast cancer. *J. Clin. Oncol.* **24,** 4611–4619 (2006).
22. Harris, L. N. *et al.* Predictors of resistance to preoperative trastuzumab and vinorelbine for HER2-positive early breast cancer. *Clin. Cancer Res.* **13,** 1198–1207 (2007).
23. Dressman, H. K. *et al.* Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy. *Clin. Cancer Res.* **12,** 819–826 (2006).
24. Potti, A. *et al.* Genomic signatures to guide the use of chemotherapeutics. *Nature Med.* **12,** 1294–1300 (2006).
25. Staunton, J. E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA* **98,** 10787–10792 (2001).
26. Lee, J. K. *et al.* A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc. Natl Acad. Sci. USA* **104,** 13086–13091 (2007).
27. Coombes, K. R., Wang, J. & Baggerly, K. A. Microarrays: retracing steps. *Nature Med.* **13,** 1276–1277 (2007).
28. Bonnefoi, H. *et al.* Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. *Lancet Oncol.* **8,** 1071–1078 (2007).
29. Neve, R. M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10,** 515–527 (2006).
30. Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genet.* **34,** 226–230 (2003).
31. Saal, L. H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl Acad. Sci. USA* **104,** 7564–7569 (2007).
32. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
33. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).
    In this study, a compendium of gene-expression profiles of cells subjected to known perturbations was used to find connections among small molecules that share a mechanism of action, between chemicals and physiological processes, and between drugs and diseases.
34. Hieronymus, H. *et al.* Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10,** 321–330 (2006).
35. Wei, G. *et al.* Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* **10,** 331–342 (2006).
36. Vogel, C. L. *et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* **20,** 719–726 (2002).
37. Berns, K. *et al.* A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer. *Cancer Cell* **12,** 395–402 (2007). This study used large-scale RNA-interference screens to identify oncogenic pathways that cause resistance to anticancer drugs in the clinic.
38. Nagata, Y. *et al.* PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. *Cancer Cell* **6,** 117–127 (2004).
39. Swanton, C. *et al.* Regulators of mitotic arrest and ceramide metabolism are determinants of sensitivity to paclitaxel and other chemotherapeutic drugs. *Cancer Cell* **11,** 498–512 (2007).
40. Whitehurst, A. W. *et al.* Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature* **446,** 815–819 (2007).
41. Farmer, H. *et al.* Targeting the DNA repair defect in *BRCA* mutant cells as a therapeutic strategy. *Nature* **434,** 917–921 (2005).
42. Bryant, H. E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434,** 913–917 (2005).
43. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnol.* **24,** 1151–1161 (2006).
44. Sparano, J. A. TAILORx: trial assigning individualized options for treatment (Rx). *Clin. Breast Cancer* **7,** 347–350 (2006).
45. Mook, S. *et al.* Individualization of therapy using MammaPrint: from development to the MINDACT Trial. *Cancer Genomics Proteomics* **4,** 147–155 (2007).
46. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347,** 1999–2009 (2002).
47. Harris, L. *et al.* American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J. Clin. Oncol.* **25,** 5287–5312 (2007).
48. Rhodes, A. *et al.* Reliability of immunohistochemical demonstration of oestrogen receptors in routine practice: interlaboratory variance in the sensitivity of detection and evaluation of scoring systems. *J. Clin. Pathol.* **53,** 125–130 (2000).
49. Ach, R. A. *et al.* Robust interlaboratory reproducibility of a gene expression signature measurement consistent with the needs of a new generation of diagnostic tools. *BMC Genomics* **8,** 148, doi:10.1186/1471-2164-8-148 (2007).
50. Hornberger, J., Cosler, L. E. & Lyman, G. H. Economic analysis of targeting chemotherapy using a 21-gene RT-PCR assay in lymph-node-negative, estrogen-receptor-positive, early-stage breast cancer. *Am. J. Manag. Care* **11,** 313–324 (2005).

Mining the plasma proteome for cancer biomarkers

Samir M. Hanash[1], Sharon J. Pitteri[1] & Vitor M. Faca[1]

Systematic searches for plasma proteins that are biological indicators, or biomarkers, for cancer are underway. The difficulties caused by the complexity of biological-fluid proteomes and tissue proteomes (which contribute proteins to plasma) and by the extensive heterogeneity among diseases, subjects and levels of sample procurement are gradually being overcome. This is being achieved through rigorous experimental design and in-depth quantitative studies. The expected outcome is the development of panels of biomarkers that will allow early detection of cancer and prediction of the probable response to therapy. Achieving these objectives requires high-quality specimens with well-matched controls, reagent resources, and an efficient process to confirm discoveries through independent validation studies.

The goal of the cancer biomarker field is to develop simple non-invasive tests that indicate cancer risk, allow early cancer detection, classify tumours so that the patient can receive the most appropriate therapy and monitor disease progression, regression and recurrence. In addition, biomarkers can be used in combination with imaging techniques to detect cancer *in vivo* and assess response to therapy (see page 580).

The dynamic nature of the circulatory system and its constituents reflects diverse physiological or pathological states, and the ease with which the blood can be sampled makes it a logical choice for biomarker applications. Blood components that provide an indication of cancer status include various cellular elements such as circulating tumour cells, cell-free DNA and RNA, proteins, peptides and metabolites. Proteins that are detectable in serum and plasma form the basis of commonly used tests to screen and monitor prostate cancer through the measurement of prostate-specific antigen. Proteins are also used to monitor response to therapy and recurrence of disease in ovarian, pancreatic and colon cancer, through measuring the antigens CA125, CA19-9 and carcinoembryonic antigen, respectively[1]. These proteins occur in serum at concentrations that are detectable by current proteomic profiling technologies. Therefore, it is reasonable to assume that there are many more cancer biomarkers present in similar concentrations in the serum and plasma that will be identified by systematic searches.

Thousands of publications have explored the potential use of individual proteins or collections of proteins as cancer biomarkers and have produced promising results. Proteins were selected for investigation as biomarkers for particular cancers because of their 'track record' as biomarkers for other disease states or other cancers or because of their function or family relationships. Predictions from genomic analyses and gene-expression data have also been used to guide research. For most of these proteins, however, their role as a cancer biomarker has not been validated in the context of a defined clinical application.

The advent of proteomic technologies has allowed the systematic interrogation of complex proteomes and the identification of differentially expressed proteins in cells, tissues and body fluids. This has heightened interest in the application of proteomics to the identification of cancer biomarkers found in the blood. In this review, we explore the progress that has been made so far, particularly in the application of mass spectrometry to biomarker discovery, as well as the pitfalls that have been encountered, and the opportunities and challenges that lie ahead in the development of biomarker panels.

Current proteomic technologies

Mass spectrometry was initially used as a tool to identify and characterize isolated proteins and to profile the mass of components in clinical samples by using matrix-assisted laser desorption/ionization (MALDI) techniques. It has evolved into a technique for interrogating complex proteomes by matching mass spectra to sequence databases to identify proteins[2]. However, even with the substantial improvements in sensitivity and mass accuracy over the past decade, the complexities of tissue and biological-fluid proteomes far exceed the current capability of mass spectrometry to resolve the individual protein and peptide constituents fully in a single analysis.

Current strategies to achieve in-depth coverage require sample fractionation followed by separate analyses of individual fractions[3] or capture of protein or peptide subsets such as glycosylated proteins or phosphopeptides[4] (Fig. 1). The trade-off for in-depth analysis is limited throughput with extensive fractionation or limited coverage of the proteome if a protein subset is singled out for analysis. Quantitative analysis of protein and peptide constituents can be achieved by isotopic labelling of proteins and peptides or by label-free quantification of derived mass spectra.

Current trends in mass spectrometry include peptide identification from accurate mass[5], increased emphasis on the characterization of post-translational modifications[2], development of ways to increase peak capacity[6] and development of high-throughput strategies for validation studies[7]. Various microarray formats — in which protein-capture agents (such as antibodies or aptamers, which are peptides, or DNA or RNA molecules that bind to a specific target molecule), recombinant or natural proteins, or cell or tissue lysates are arrayed and then interrogated with patient samples — are slowly complementing mass spectrometry as a high-throughput tool for biomarker discovery and validation[8]. However, mass spectrometry in combination with liquid chromatography will remain the proteomic workhorse until microarrays and capture agents can comprehensively interrogate complex proteomes.

[1]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA.

Developing biomarker profiles
Profiling tissues to identify potential circulating biomarkers

The abundance range of known plasma proteins spans nine orders of magnitude. Tumour-tissue-derived proteins in the circulation are probably present at the lower end of this range, particularly during the early stages of tumour development. Given that tumour-derived proteins are present at higher concentrations closer to their source, analysis of tumour cells, tissues and proximal fluids to identify differentially expressed proteins has been pursued as an alternative strategy to comprehensive profiling of plasma for biomarker discovery. Potential protein biomarkers in plasma might be deduced from proteomic and/or transcriptomic profiling, with subsequent validation studies of candidate biomarkers performed in serum or plasma. It is largely unknown whether there is a link between restricted expression at the tissue level and the occurrence of candidate biomarkers at detectable levels in serum in a manner that matches their expected disease-specificity profile based
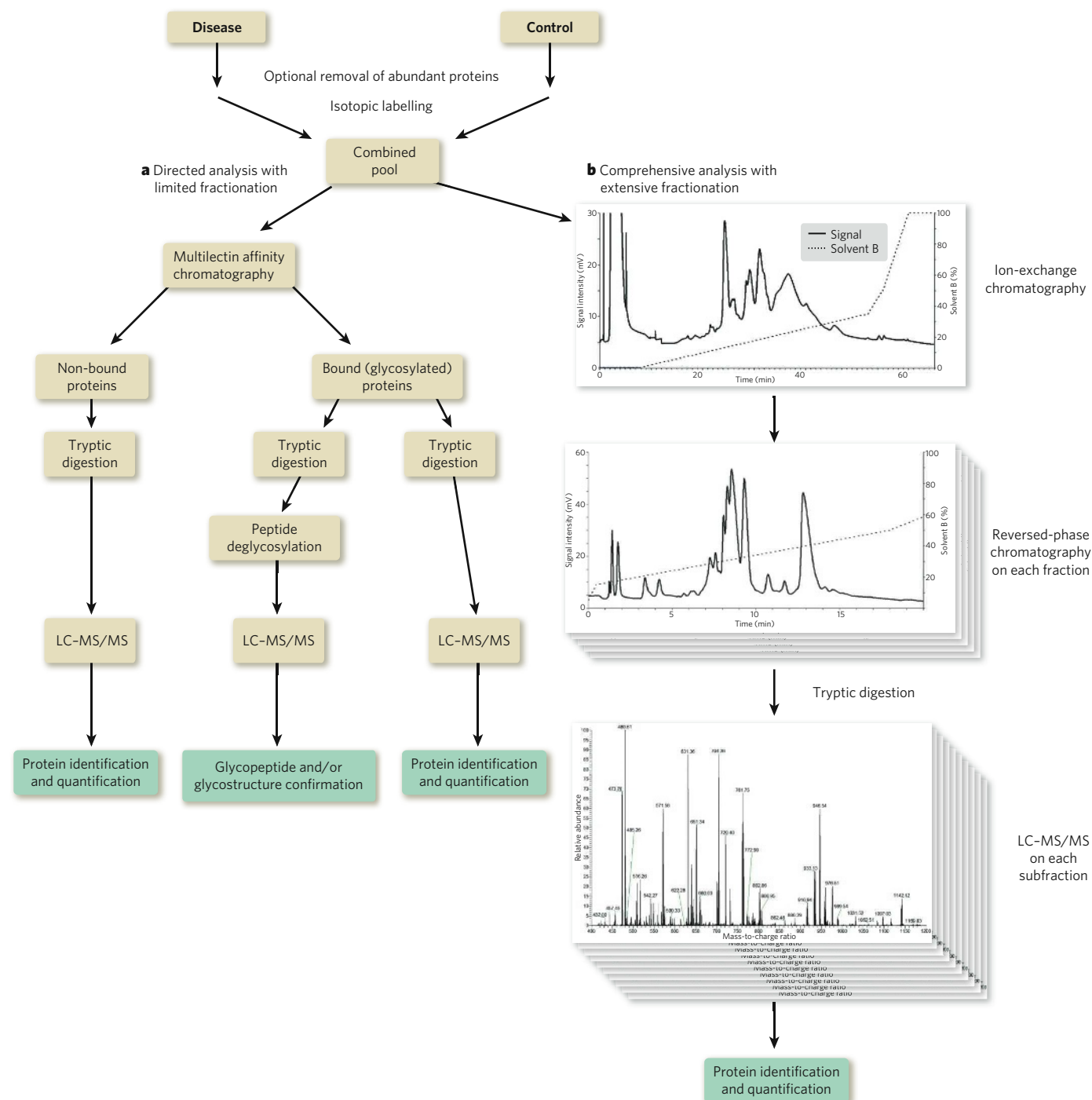
Figure 1 | In-depth analysis of plasma proteins. To overcome the limited dynamic range of mass-spectrometry detectors and the effect of undersampling, approaches have been developed to allow identification of low-abundance proteins. Two basic approaches are illustrated. The first approach is to select a subset of plasma proteins for detailed analysis, as exemplified here by affinity chromatography using multiple lectins to capture glycoproteins (**a**). The second approach is to subject plasma proteins to extensive fractionation (by ion-exchange chromatography and reversed-phase chromatography) before analysis of individual fractions (**b**). With both approaches, low-abundance proteins are discovered as a result of the reduced complexity of individual fractions subjected to analysis by liquid chromatography–tandem mass spectrometry (LC–MS/MS). (Panel **a** adapted, with permission, from ref. 49. Panel **b**, adapted, with permission, from ref. 3.)

on tissue analysis. This is because few of the candidate biomarkers that have emerged from tissue analysis have been subsequently validated as useful biomarkers in serum or plasma. Factors that may contribute to discrepancies between tissue and plasma findings include heterogeneity within tumours not represented in samples subjected to profiling, difficulty in defining and obtaining adequate controls, and the release into and clearance of proteins from the circulation in a manner that is not predictable from tissue analysis. For example, a protein may be differentially expressed at the tumour level compared with unaffected tissues, but the resultant increment in circulating levels of the protein may be negligible owing to the greater mass of unaffected tissues releasing the protein into the circulation. Alternatively, levels of the protein in tumour tissue may be unchanged compared with unaffected tissues, but because of altered processing, increased turnover and cell breakdown, the protein may occur at increased concentrations in the circulation[9]. However, the evidence available is anecdotal, and there is a need to collect data to rigorously determine the relationship between tissue expression, release into and clearance from the circulation, and the factors involved.

Proximal biological fluids

Various types of fluid and effusion offer a window onto the proteins from tumour tissue that may be released into extracellular fluids through secretion or cell and tissue breakdown[10]. Such fluids include breast ductal fluid, which can be obtained from nipple aspiration and ductal lavage for breast cancer, pleural effusions in the case of intrathoracic tumours, ascitic fluid for intra-abdominal tumours, and cerebrospinal fluid for central nervous system tumours. Numerous studies have focused on nipple aspirates and ductal lavage for identifying potential breast-cancer candidates. These studies present several challenges. The procedures for obtaining fluid are difficult to standardize. Controls may not be adequate even when obtained from a contralateral, presumably unaffected, breast, owing to disparities in the amount of fluid obtained and the occurrence of microhaemorrhages, for example. The amount of fluid obtained and the presence of haemorrhages might also depend on confounding inflammatory conditions. The protein constituents of such biological fluids can be heterogeneous as a result of varying cell and tissue admixture and breakdown, making it difficult to sort out potential biomarkers from extraneous sources.

A review of data for tissue, pancreatic juice and blood profiling for pancreatic-cancer biomarkers from published studies reveals some agreement between the three sets on the differential expression of relatively abundant serum proteins (Table 1, top). However, these proteins as a group reflect inflammatory changes, because they are also increased in abundance during pancreatitis and other inflammatory conditions. Proteomic analyses of pancreatic tissue and pancreatic juice from individuals with or without cancer yield inconsistent results for the regulated proteins, which makes it difficult to determine which protein changes are the most relevant as biomarkers for diagnosing cancer (Table 1, bottom). However, discrepancies may be expected for secreted proteins, which are more likely to be found in pancreatic juice than tissue.

Tumour-cell populations

Proteomic profiling of tumour cell lines and dysplastic or malignant cells isolated from effusions or from disaggregated tumour tissue has the advantage of allowing separate analysis of individual subcellular compartments that are particularly relevant to protein release into the plasma, notably the secretome and the cell-surface proteomes (Fig. 2). Identification of informative protein subsets from these sources is well within the capability of proteomic technologies.

As part of a search for potential diagnostic biomarkers for lung cancer, apical surface fluids from aberrantly differentiated squamous metaplastic human tracheobronchial epithelial cells in culture and from normal counterparts were compared[11]. From this, 22 proteins were identified that were associated with metaplasia, including S100A8, S100A9, annexin I and annexin II. In another study of breast-cancer cell lines,

proteomic analysis of conditioned media provided a rich data set of secreted proteins with established roles in breast-cancer development and identified additional potential biomarkers[12]. In all, large data sets of secreted proteins and candidate biomarkers have emerged from studies of cell populations, and these will probably contribute plasma biomarkers once related assays have been developed and validation studies have been performed.

Depth of analysis of the plasma proteome

Given the low abundance in serum and plasma of known cancer biomarkers, the issue is whether current proteomic technologies provide sufficient depth of analysis for biomarker discovery. Several recent studies have investigated the extent to which proteomic technologies can unravel the complexity of the plasma proteome.

The Human Proteome Organisation completed a collaborative study to characterize the human serum and plasma proteomes[13]. Diverse methods and instruments were used to compare and integrate tandem mass spectrometry (MS/MS) data from aliquots of pooled serum and plasma from healthy subjects. Initial data integration resulted in 3,020 proteins with two or more peptides identified. The study applied rigorous statistical approaches to take into account the length of coding regions in genes and multiple hypothesis-testing techniques. On that basis, a reduced set of 889 proteins was identified with a confidence level of at least 95%. Of relevance are the high-confidence identifications of protein matches to novel exons, some of which were localized in alternatively spliced forms of known plasma proteins and others in previously non-annotated gene sequences. The concentrations of some of the identified proteins are well within the range of known cancer biomarkers, and the list of high-confidence protein identifications includes proteins such as CA125, which clearly points to the ability to identify low-abundance proteins.

Another study tested the assumption that proteins derived from a variety of tissues could be detected in plasma by using current proteomic technologies[14]. Results obtained indicate that such tissue-derived proteins were detectable in plasma by direct mass-spectrometric analysis of captured glycopeptides, thus providing support for plasma-based protein biomarker discovery. Further evidence that substantial depth of analysis is currently achievable with fractionation comes from a plasma study following a two-dimensional protein-fractionation schema that resulted in high-confidence identification of 1,662 proteins and included protein isoforms that differed in their chromatographic mobility[3]. Interestingly, further increase in depth of analysis was accomplished simply by repeat analysis of aliquots from the same set of plasma fractions, resulting in overall identification of 2,254 proteins. Repeat analysis overcomes partial sampling of low-abundance peptides in single liquid-chromatography–MS/MS analyses. Thus, current schemes for analysis of a plasma protein subset (for example, glycoproteins or fractionated plasma) allow sufficient depth of analysis to identify circulating cancer biomarkers.

The plasma proteome for cancer biomarker identification

Apart from the technological challenge presented by the range of protein concentrations, the profiling of serum or plasma for cancer biomarker identification is difficult in numerous other ways. Perhaps the most important are the sources of variability that contribute to a false discovery and therefore have to be taken into consideration as part of the experimental design and data interpretation (Box 1). Artefacts or nonspecific disease-associated changes in plasma proteins need to be distinguished from potentially specific biomarkers. This is highly pertinent because cancer-serum protein profiling by MALDI mass spectrometry has uncovered mass profiles for previously unidentified proteins that were proposed to be diagnostic for several common types of cancer but whose validity was subsequently questioned[15]. It is doubtful whether distinct mass peaks observed truly represent cancer-specific biomarkers; instead, they are more likely to result from nonspecific changes, given the limited sensitivity of the approach used and, therefore, the likely correspondence of the mass peaks to relatively

573

Table 1 | Comparison of published findings from proteomic analysis of pancreatic tissue, pancreatic juice and serum or plasma

Gene symbol	Protein name	Pancreatic tissues		Pancreatic juice		Plasma or serum		
		Ratio pancreatic cancer to normal[44]	Ratio chronic pancreatitis to normal[45]	Ratio pancreatic cancer to normal[46]	Ratio chronic pancreatitis to normal[46]	Ratio post-resection to pre-resection[47]	Ratio recurrent to non-recurrent[47]	Ratio pancreatic cancer to chronic pancreatitis[48]
Proteins upregulated in serum or plasma and identified in pancreatic tissue or juice								
A1BG	α_{1B}-Glycoprotein	–	–	↔	↑	↑	–	–
A2M	α_2-Macroglobulin	–	↑	↔	↑	↑	↑	↓
ALB	Albumin	↑	↑	↑	↑	–	↑	–
AMBP	α_1-Microglobulin/bikunin	↑	–	–	–	–	–	↓
B2M	β_2-Microglobulin	↑	↑	↑	–	↑	–	–
C3	Complement component C3	–	↑	↑	↔	↑	↑	↑
C9	Complement component C9	↓	–	–	–	–	–	↑
CFB	Complement factor B	↑	↑	–	–	–	–	↑
CFH	Complement factor H	–	↑	–	–	↑	–	–
HPX	Haemopexin	↑	↑	–	–	↑	–	↓
IGHM	Immunoglobulin μ-chain constant region	↑	–	↑	↑	–	–	–
PLG	Plasminogen	–	↑	–	↑	–	↑	↑
Proteins upregulated in tissue or pancreatic juice with an unknown status in plasma and serum								
CTRB1	Chymotrypsinogen B1	↓	–	↑	↑	–	–	–
ELA3B	Elastase 3B	↓	–	↑	↔	–	–	–
GP2	Pancreatic secretory granule membrane major glycoprotein GP2	↓	–	↑	↔	–	–	–
PRSS1	Trypsin I	↓	–	↑	↑	–	–	–
PRSS3	Trypsin III	↓	–	↑	–	–	–	–
REG1A	Lithostathine 1α	↓	↑	↑	↔	–	–	–
FGB	Fibrinogen β-chain	↑	↑	↑	↑	–	–	–
FGG	Fibrinogen γ-chain	↑	↑	↑	↔	–	–	–
IGHA1	Immunoglobulin α1-chain constant region	↑	–	↑	↑	–	–	–
PSAP	Prosaposin	↑	–	↓	–	–	–	–

Data are taken from references listed in each column heading. ↑, upregulated; ↓, downregulated; ↔, no change; –, protein not detected.

abundant serum proteins. An informative study in this regard is an antibody-microarray-based quantitative analysis of abundant serum proteins in 80 samples from newly diagnosed subjects with lung cancer, healthy controls and subjects with chronic obstructive pulmonary disease[16]. Seven of the 84 arrayed antibodies, which included those that targeted C-reactive protein, serum amyloid A, mucin 1 and α_1-antitrypsin, produced a significant difference ($P<0.01$) between patients with lung cancer and controls. Leave-one-out cross-validation was used to construct diagonal linear discriminant analysis classifiers. At a cutoff where all 56 of the non-tumour samples were correctly classified, 15 of 24 lung-tumour sera were also correctly classified. These findings suggest that profiling of relatively abundant serum proteins may yield distinct differences between cancers and controls. However, such differences may not be specific for cancer or a cancer type, and their utility as part of a strategy for detecting cancer has yet to be determined.

Types of biomarker
Biomarkers for response to therapy
Biomarkers do not need to be cancer-specific to be useful; for example, certain proteins may help to predict response to therapy or aid in the monitoring of disease progression. Although biomarkers such as CA125 or CA19-9 may have limited value for early detection because they are present at increased levels in benign disease, they are valuable for monitoring cancer progression and response to therapy. Furthermore, as cancer is increasingly defined by dysregulated pathways, relevant biomarkers may cut across tumour types without showing tissue specificity. Moreover, although biomarker applications for screening

healthy populations are highly demanding with respect to specificity, other applications may be less demanding.

Encouraging findings have been reported in the search for proteomic signatures that predict response to targeted therapies. In one study, protein-expression profiles of lung adenocarcinoma tissues from patients treated with gefitinib were analysed[17]. Nine proteins that distinguished responders from non-responders were identified. The predictive performance of the nine proteins was validated by examining an additional group of responders and non-responders. In addition, 12 other proteins constituted a signature that distinguished tumours on the basis of the mutation status of the epidermal growth-factor receptor (EGFR) gene. In an independent serum-based study by a different group, MALDI mass-spectrometric analysis of pretreatment serum was used to identify patients who would probably benefit from treatment with inhibitors of EGFR tyrosine-kinase activity[18]. An algorithm was developed to predict treatment outcomes on a training set of 139 patients from three cohorts, which was then tested on two independent validation cohorts of 67 and 96 patients treated with inhibitors of EGFR tyrosine-kinase activity and on three control cohorts of patients who were not treated with these drugs. For both validation cohorts, the classifier identified patients who showed improved outcomes after drug treatment. Median survival of patients in the predicted favourable survival and poor survival groups was 207 days versus 92 days, respectively, for one validation group, and 306 days versus 107 days for the other. Although substantial efforts were made to assess technical reproducibility, the identity of the proteins that predicted the response was not determined. Therefore, the biological significance of the informative proteins is unclear, and assays for their independent validation cannot be developed at present.
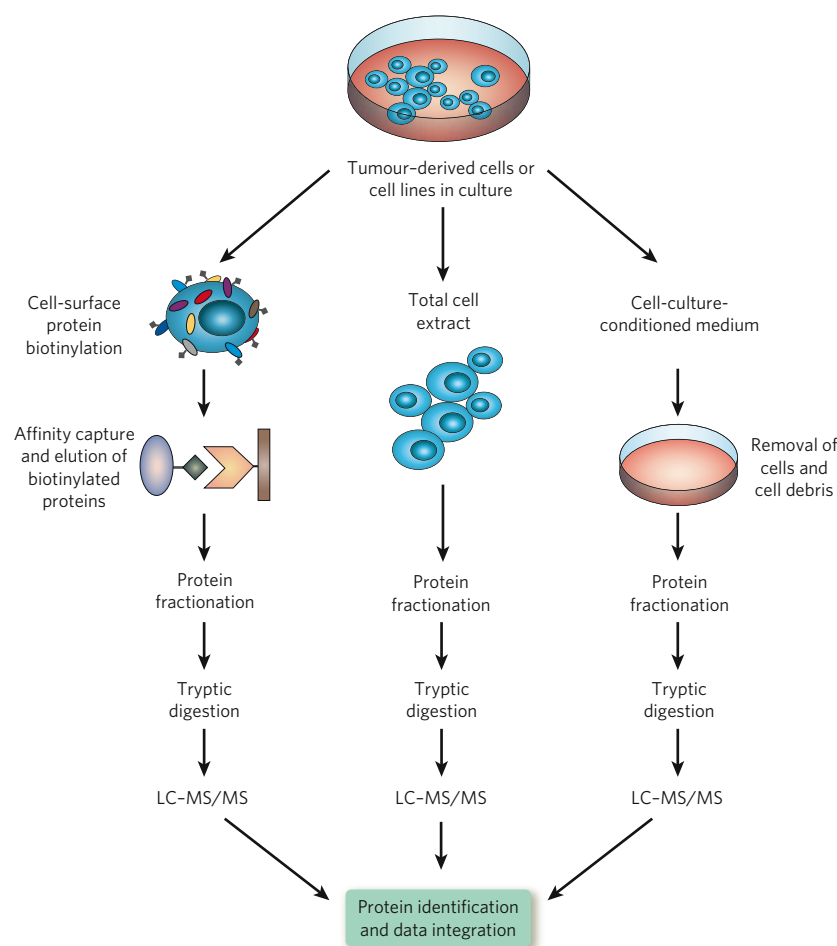
Figure 2 | Searching for biomarkers through analysing the cellular proteome. A comprehensive proteomic profile of tumour cells or cell lines in culture can be obtained by analysing subcellular 'compartments': namely, the cell surface and the secretome. Cell-surface proteins (left) are obtained by labelling proteins on the outside of intact cells with a biotinylated tag and then capturing these proteins by affinity chromatography (through the affinity of biotin for avidin). Secreted proteins (right) are obtained directly from cell-conditioned medium after cells and cell debris have been removed carefully. In parallel with the total cell extract (centre), these two subcellular sets of proteins are then fractionated (by reversed-phase chromatography), and individual fractions are subjected to tryptic digestion followed by LC–MS/MS, for identification of proteins. With such an approach, an in-depth analysis of the cellular proteome is obtained and subsequently mined to identify potential biomarkers.

Post-translational modifications as biomarkers

Post-translational modifications of proteins are important in many biological processes, and post-translational changes have relevance to disease and cancer. The study of post-translational modifications as a source of biomarkers in cancer is still at a relatively early stage. Cleavage products of tumour-derived proteins have been proposed as potential cancer biomarkers. Higher levels of calreticulin and protein disulphide isomerase A3 fragments have been found in the serum of patients with hepatocellular carcinoma than in controls[9]. Cleavage products of common serum proteins have also been proposed to provide a peptide signature for cancer[19]. However, the mechanism that confers specificity for cancer or cancer types on such products remains to be demonstrated.

Altered protein glycosylation in cancer is another source of potential cancer biomarkers[20]. Identification of glycosylated proteins relies on various glyco-capture strategies, a means of glycosylated-protein sub-selection by affinity chromatography. Cancer-specific glycans have been identified as potential biomarkers for prostate cancer[21]. Increased fucosylation of serum ribonuclease 1 and altered glycosylation of other proteins have been reported in pancreatic cancer[22,23]. Studies of EGFR glycosylation illustrate the depth of analysis that can be obtained by targeted approaches[4]. Ten extracellular-domain N-linked glycosylation sites were identified, and the main glycoforms postulated to be present at each site were quantified in cancer cells, with partial characterization of EGFR glycosylated peptides in serum (Fig. 3). In the coming years, the study of post-translational modifications will undoubtedly benefit from emerging mass-spectrometric technologies such as electron-transfer dissociation[24], which allows labile modifications to remain intact while obtaining peptide sequence information, enabling the study of modifications such as glycosylation and phosphorylation.

The challenge of early cancer detection

For most tumour types, the earliest stages of tumour development remain poorly understood, and the tissues involved are ill defined or inaccessible, thus impeding early detection. Detection of cancer at the tumour-inception stage by measuring circulating protein biomarkers may not be within reach. However, the prospects are less gloomy for the development of biomarkers that allow detection of cancer at a stage early enough for it to be cured. A better understanding of the rate of release of proteins from tumours at early stages and the extent to which a field effect occurs (that is, increased release of the same proteins from surrounding cell populations), resulting in an amplification process and a detectable signal in blood, is needed.

A promising approach for early cancer detection is to harness the immune response directed against tumour antigens. There is increasing evidence of an immune response to cancer in humans, as demonstrated in part by the identification of autoantibodies specific for a number of intracellular and cell-surface antigens detectable in sera from patients with different cancer types. Clearly, tumours can still develop in the presence of this antibody-mediated immune response. However, the identification of a panel of antigenic biomarkers that are tumour-specific and that elicit immunoreactivity early in tumour development and at a high frequency would provide an effective strategy for cancer screening. Harnessing the immune response to identify cancer biomarkers is attractive because the immune system carries out a biological amplification, which generates a detectable response to antigenic tumour proteins. Several strategies have been pursued to determine the repertoire of antibodies in the sera of patients with cancer. A large number of antigenic targets have been identified by screening expression libraries with patient sera. However, most of the identified antigens elicited antibodies in a relatively small proportion of patients. Phage display has also been used to find antigen candidates[25,26].

Proteomics is increasingly relied on to identify tumour antigens that induce an antibody response. Recombinant protein microarrays have been used to screen for autoantibodies in ovarian cancer[27]. Sera from 30 patients with ovarian cancer and 30 healthy individuals were used to probe microarrays containing 5,005 human proteins for immunoglobulin reactivity. Ninety-four antigens were identified that showed more reactivity with sera from cancer patients than control sera. Reactive antigens tested using tissue microarrays were found to have increased expression in ovarian-cancer tissue relative to controls, which provides support for their immune recognition as aberrantly expressed proteins. Other proteomic approaches allow analysis of proteins and peptides in their post-translationally modified states, as they occur in cancer cells. Modifications, such as glycosylation, may be immunogenic, so there is merit in using approaches that preserve such epitopes. Mass spectrometry was used to identify tissue proteins that form a complex with circulating autoantibodies in ovarian cancer[28]. Tumour proteins immunoprecipitated by autoantibodies from sera from patients with cancer but not from cancer-free controls were investigated. A subset of antigens that was processed and then presented by major histocompatibility complex class I molecules on the surface of ovarian cancer cells was identified, and this subset is therefore common to both humoral and cell-mediated immunity and, consequently, has diagnostic and immunotherapeutic potential.

In a proteomic study of lung cancer that searched for serum immunoglobulins that recognize lung-cancer antigens, sera from 60% of patients with lung adenocarcinoma (including early-stage disease) but none of the cancer-free controls showed reactivity against proteins identified as glycosylated annexin I and annexin II (ref. 29). In subsequent blinded validation studies, annexin-specific antibodies were detected in sera collected a year before a diagnosis of lung cancer was made. Annexin-specific autoantibodies, together with autoantibodies specific for the proteins PGP9.5 and 14-3-3-θ, gave a sensitivity of 55% at 95% specificity in discriminating lung cancer at the preclinical stage from matched controls[30]. Microarrays containing the repertoire of natural proteins expressed in tumour cells have the potential to substantially accelerate the pace of discovery of tumour antigens and could provide a molecular signature for immune responses in different types of cancer[31].

Validation strategies for discovered protein biomarkers

The path from discovery to approval by the US Food and Drug Administration is arduous for any biomarker, and proteomics-derived cancer biomarkers are no exception. The biomarker validation process has been divided into five phases[32]. Assay development for validation studies remains a major hurdle, and the number of samples needed for validation increases as the biomarker advances through the phases; hence

the need for high-throughput assays. The most relied-on approach for validation remains the sandwich enzyme-linked immunosorbent assay (ELISA), which is highly specific because it uses a pair of antibodies against the targeted molecule. New developments in nanotechnology allow the use of antibodies coupled to a wide variety of new fluorescent molecules, such as quantum dots, enabling sensitive detection up to the point of visualizing single molecules[33] and multiple molecules simultaneously[34]. These techniques are also suitable for the application of biomarkers to tumour imaging (see page 580). The limits of detection have also been increased using amplification strategies that take advantage of PCR with rolling circle amplification[35] or proximity ligation[36]. Technological developments, such as surface plasmon resonance[37], are also improving the limits of detection for complexes, and nanomechanical resonators can accurately measure the mass of molecules[38]. However, the limiting factor for all these validation approaches is the availability of antibodies with the required affinity and specificity for the targets in need of validation.

Approaches that do not require antibodies continue to be explored. An attractive alternative is the use of multiple reaction monitoring (MRM) by mass spectrometry to monitor fragments of specific peptides, for the direct quantification of the corresponding proteins in serum[39]. In general, target peptides labelled with stable isotopes are used as internal standards, allowing very accurate measurements. One of the advantages of this approach is the multiplexing feature, allowing rapid characterization of several clinically relevant plasma proteins simultaneously and in a high-throughput mode. This approach still has limitations in terms of sensitivity, with detection limits below 1 µg ml^{-1} in crude plasma, and it is applicable only to previously well-characterized peptides.

Peptide[40], RNA or DNA[41] aptamers also hold some potential to overcome limited availability of antibodies for diagnostics.

What lies ahead

Although substantial depth of analysis is currently feasible, most of the published proteomics-based studies have limited sensitivity and therefore were unable to detect potential biomarkers that occur in low abundance. Given this lack of success, there is a need to take advantage of the full capability of mass spectrometry for in-depth quantitative profiling. This requires an elaborate set-up with substantial mass-spectrometric and informatics resources. It also requires a reproducible platform to yield quantitative data and results for protein and peptide identifications and, preferably, to define post-translational modifications.

Experimental design has a crucial role in a successful biomarker search. It begins with the choice of specimens and the adequacy of matching between cases and controls to avoid bias. The availability of high-quality, matched specimens has been limited, and investigations are often compromised as a result. Greater effort is being made to provide investigators with the most informative specimens. This is exemplified by the development of specimen resources that are available to investigators as part of the National Cancer Institute Early Detection Research Network (http://www.edrn.nci.nih.gov). Specimens are also available from various cohort studies such as the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (http://www.plco.gov.org). Although cohort-related specimens were previously available only for validation studies, there is increased appreciation of the value of such specimens for discovery. For example, proposals that take advantage of the availability of specimens collected as part of the Women's Health Initiative (http://www.nhlbi.nih.gov/whi) are now welcomed. This initiative is engaged in studies and longitudinal follow-up of over 160,000 women. Cohort specimens collected before the onset of symptoms from subjects who later developed cancer are particularly useful for risk assessment and for studies of biomarkers for early detection.

The challenge of discovering cancer biomarkers necessitates integrated and cooperative efforts. Most studies so far have been undertaken by individual investigators who have focused on analysis of a single type of specimen for biomarker investigations. There is a need to implement strategies to share and integrate data from analyses of various sources,

including tumour-cell populations, tumour tissue, proximal fluids and peripheral blood, to select the most promising biomarkers. Such a goal is best achieved through organized, cooperative efforts that develop and follow road maps for biomarker discovery and validation and that engage multidisciplinary teams.

Previously, most biomarker discovery studies were not driven by a particular clinical application. Instead, they tended to be of a type in which disease specimens were compared with controls without consideration of the requirements of a clinical application. However, there is merit in defining the application for which biomarkers are being sought to ensure that the most relevant specimens for discovery and initial validation are used. For predictive biomarkers, this could be accomplished by embedding biomarker discovery and validation into clinical trials.

Although biomarker discovery is likely to entail a painstaking process of systematically searching the proteome for potential candidates, the validation process is more direct. However, validation studies have obstacles of their own. Most noteworthy is the paucity of affinity-capture

agents for detailed investigations of candidate biomarkers. Mass-spectrometry-based strategies are currently in development for biomarker validation through analysis of representative peptides. However, there remains a pressing need for affinity-capture agents, notably high-quality antibodies. These agents allow investigations of candidate biomarkers that are not feasible with mass spectrometry alone. Examples of such investigations include elucidation of the tissue distribution and subcellular localization of candidate biomarkers, their occurrence as part of a complex, and the development of high-throughput assays for their measurement using a variety of assay formats. A noteworthy development is the emerging partnership between the public and private sectors for developing broad sets of affinity-capture agents such as the Human Protein Atlas project through the support of the Knut & Alice Wallenberg foundation (http://www.proteinatlas.org), and to develop antibodies specific for ovarian-cancer and pancreatic-cancer candidate biomarkers through support from the Canary Foundation and the Lustgarten Foundation (http://www.canaryfoundation.org; http://www.lustgarten.org).
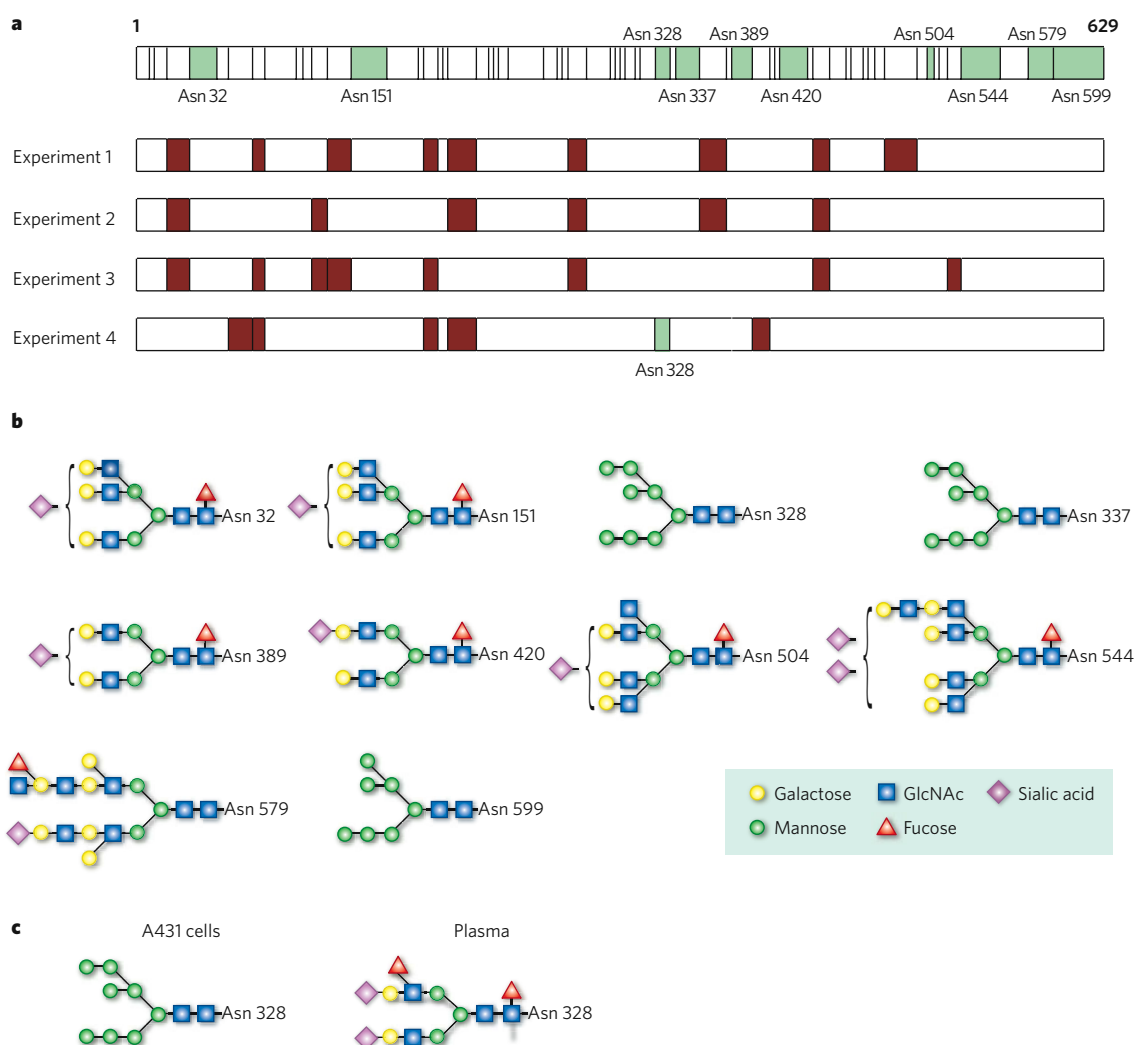
Figure 3 | Identification of EGFR glycoforms. a, The extracellular-domain sequence of EGFR (629 amino acids) is shown at the top. Bars indicate peptides that are predicted to be generated by tryptic digestion. Green boxes indicate peptides with known N-linked glycosylation sites. Red boxes indicate peptides identified in a series of four LC–MS/MS experiments to identify proteins in reference pools of plasma from healthy subjects and patients with cancer. A manual inspection for known glycosylated peptides from EGFR was then carried out, leading to the identification of the peptide containing the Asn-328 glycosylation (as indicated by the green box in experiment 4). **b,** EGFR from the A431 cell line was subjected to mass-spectrometric analysis. Ten N-linked glycosylation sites were identified, and the main glycan proposed to be at each site is shown[4]. Curly brackets indicate that the monomer (sialic acid) is present in one of these positions (or two in the case of the glycan attached to Asn 544). **c,** The proposed glycan structure for one EGFR peptide differs from that of the corresponding peptide identified in A431 cells. This structure was determined following LC–MS/MS analysis of human plasma from a mixture of control individuals and patients with lung cancer (S. L. Wu and B. L. Karger, personal communication). GlcNAc, N-acetylglucosamine.

577

Towards a rational design of biomarker panels

It is clear that single biomarkers are unlikely to provide the sensitivity and specificity needed for most applications given the substantial heterogeneity among cancers. It is unrealistic to expect that a single biomarker will provide information about tissue type and malignant transformation throughout the various stages of tumour development and progression. Therefore, panels of biomarkers are needed. However, assembling and validating such a panel is quite a challenge. A candidate biomarker may be dismissed if in initial studies it provides modest sensitivity. However, such a biomarker may be valuable if it is informative with respect to a particular subset of subjects whose disease subtype is not obvious.

Rational approaches to the development of biomarker panels are worthy of consideration. An iterative approach to assembling a panel, for example for cancer screening, would be to eliminate subsets of cases for which available biomarkers perform well and concentrate instead on those subsets for which new biomarkers are needed. This iterative process would continue until a panel is assembled that is applicable across all subjects for whom the screening test is intended.

Another approach would be to define expectations for tissue specificity separate from considerations of malignancy. In that scenario, a biomarker(s) may reflect pathology in a particular tissue or organ type but may not be particularly specific for cancer. However, when combined with biomarkers that inform with respect to cancer but on their own do not inform with respect to organ or tissue of origin, the combination of such sets of biomarkers would be more fully informative.

The development of circulating biomarkers and autoantibodies for diagnostic and predictive applications has followed separate paths. However, it would be useful to consider integrating the development of circulating tumour-derived proteins with an understanding of the contribution of the corresponding autoantibodies. Levels of a circulating tumour biomarker may be reduced as a result of the formation of complexes with autoantibodies, contributing to a decrease in biomarker sensitivity. Alternatively, serum from a subject with cancer may not show any evidence of an autoantibody because of excess corresponding tumour antigen and a relatively weak immune response. Therefore, assays for autoantibodies and assays for circulating tumour biomarkers may be complementary.

Yet another approach to consider is the use of animal models for biomarker discovery. Genomic analysis of human and mouse cancers has revealed significant agreement between chromosomal aberrations and gene-expression profiles, and has established cross-species analyses as an effective filter in the identification of genes and loci embedded within complex cancer genomes[42]. Genetically engineered mouse models provide defined stages of tumour development, homogenized breeding and environmental conditions, and standardized blood sampling, thereby reducing biological and non-biological heterogeneity. The concept that plasma from genetically engineered mouse models of cancer may yield biomarkers is attractive. Interestingly, the tumour-antigen repertoire identified in tumour-bearing mice has been found to recapitulate the repertoire observed in humans[43].

The power of proteomics for the development of cancer biomarkers has yet to be taken full advantage of. Given the substantial challenges involved, the field is best served through cooperative efforts to implement appropriate experimental designs and to use the most informative matched specimens. Implementation of approaches that integrate findings from studies of tumour cells and tissues, proximal fluids and peripheral blood is likely to yield a rich set of circulating cancer biomarkers for cancer molecular diagnostic tests. ■

1. Ludwig, J. A. & Weinstein, J. N. Biomarkers in cancer staging, prognosis and treatment selection. *Nature Rev. Cancer* **5,** 845–856 (2005).
2. Cox, J. & Mann, M. Is proteomics the new genomics? *Cell* **130,** 395–398 (2007).
   This is a review of current mass-spectrometry-based proteomic technologies. It explores the relationship between genomics and proteomics and the outlook for protein-expression analysis.
3. Faca, V. *et al.* Contribution of protein fractionation to depth of analysis of the serum and plasma proteomes. *J. Proteome Res.* **6,** 3558–3565 (2007).
4. Wu, S. L. *et al.* Dynamic profiling of the post-translational modifications and interaction partners of epidermal growth factor receptor signaling after stimulation by epidermal growth factor using Extended Range Proteomic Analysis (ERPA). *Mol. Cell. Proteomics* **5,** 1610–1627 (2006).
5. Liu, T., Belov, M. E., Jaitly, N., Qian, W. J. & Smith, R. D. Accurate mass measurements in proteomics. *Chem. Rev.* **107,** 3621–3653 (2007).
6. Liu, X. *et al.* Mapping the human plasma proteome by SCX–LC–IMS–MS. *J. Am. Soc. Mass Spectrom.* **18,** 1249–1264 (2007).
7. Stahl-Zeng, J., Lange, V., Ossola, R., Aebersold, R. & Domon, B. High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Mol. Cell. Proteomics* **6,** 1809–1817 (2007).
8. Kingsmore, S. F. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nature Rev. Drug Discov.* **5,** 310–320 (2006).
9. Chignard, N. *et al.* Cleavage of endoplasmic reticulum proteins in hepatocellular carcinoma: detection of generated fragments in patient sera. *Gastroenterology* **130,** 2010–2022 (2006).
10. Hu, S., Loo, J. A. & Wong, D. T. Human body fluid proteome analysis. *Proteomics* **6,** 6326–6353 (2006).
11. Kim, S. W. *et al.* Proteomics-based identification of proteins secreted in apical surface fluid of squamous metaplastic human tracheobronchial epithelial cells cultured by three-dimensional organotypic air–liquid interface method. *Cancer Res.* **67,** 6565–6573 (2007).
12. Kulasingam, V. & Diamandis, E. P. Proteomic analysis of conditioned media from three breast cancer cell lines: a mine for biomarkers and therapeutic targets. *Mol. Cell. Proteomics* **6,** 1997–2011 (2007).
13. States, D. J. *et al.* Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nature Biotechnol.* **24,** 333–338 (2006).
    This paper describes the first large-scale collaborative study to characterize the human serum and plasma proteomes.
14. Zhang, H. *et al.* Mass spectrometric detection of tissue proteins in plasma. *Mol. Cell. Proteomics* **6,** 64–71 (2007).
15. Ransohoff, D. F. Bias as a threat to the validity of cancer molecular-marker research. *Nature Rev. Cancer* **5,** 142–149 (2005).
    This review outlines the experimental pitfalls of cancer biomarker research and suggests ways to avoid bias in clinical research.
16. Gao, W. *et al.* Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. *BMC Cancer* **5,** 110, doi:10.1186/1471-2407-5-110 (2005).
17. Okano, T. *et al.* Proteomic signature corresponding to the response to gefitinib (Iressa, ZD1839), an epidermal growth factor receptor tyrosine kinase inhibitor in lung adenocarcinoma. *Clin. Cancer Res.* **13,** 799–805 (2007).
18. Taguchi, F. *et al.* Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. *J. Natl Cancer Inst.* **99,** 838–846 (2007).
19. Villanueva, J. *et al.* Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age. *Mol. Cell. Proteomics* **5,** 1840–1852 (2006).
20. Kirmiz, C. *et al.* A serum glycomics approach to breast cancer biomarkers. *Mol. Cell. Proteomics* **6,** 43–55 (2007).
21. Kyselova, Z. *et al.* Alterations in the serum glycome due to metastatic prostate cancer. *J. Proteome Res.* **6,** 1822–1832 (2007).
22. Barrabes, S. *et al.* Glycosylation of serum ribonuclease 1 indicates a major endothelial origin and reveals an increase in core fucosylation in pancreatic cancer. *Glycobiology* **17,** 388–400 (2007).
23. Zhao, J., Qiu, W., Simeone, D. M. & Lubman, D. M. N-linked glycosylation profiling of pancreatic cancer serum using capillary liquid phase separation coupled with mass spectrometric analysis. *J. Proteome Res.* **6,** 1126–1138 (2007).
24. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl Acad. Sci. USA* **101,** 9528–9533 (2004).
25. Chatterjee, M. *et al.* Diagnostic markers of ovarian cancer by high-throughput antigen cloning and detection on arrays. *Cancer Res.* **66,** 1181–1190 (2006).
26. Wang, X. *et al.* Autoantibody signatures in prostate cancer. *N. Engl. J. Med.* **353,** 1224–1135 (2005).
27. Hudson, M. E., Pozdnyakova, I., Haines, K., Mor, G. & Snyder, M. Identification of differentially expressed proteins in ovarian cancer using high-density protein microarrays. *Proc. Natl Acad. Sci. USA* **104,** 17494–17499 (2007).
28. Philip, R. *et al.* Shared immunoproteome for ovarian cancer diagnostics and immunotherapy: potential theranostic approach to cancer. *J. Proteome Res.* **6,** 2509–2517 (2007).
29. Brichory, F. M. *et al.* An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proc. Natl Acad. Sci. USA* **98,** 9824–9829 (2001).
30. Pereira-Faca, S. *et al.* Identification of 14-3-3θ as an antigen that induces a humoral response in lung cancer. *Cancer Res.* **67,** 12000–12006 (2007).
31. Madoz-Gurpide, J., Kuick, R., Wang, H., Misek, D. E. & Hanash, S. M. Integral protein microarrays for the identification of lung cancer antigens in sera that induce a humoral immune response. *Mol. Cell. Proteomics* **7,** 268–281 (2008).
32. Yasui, Y. *et al.* A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **4,** 449–463 (2003).
33. Tada, H., Higuchi, H., Wanatabe, T. M. & Ohuchi, N. *In vivo* real-time tracking of single quantum dots conjugated with monoclonal anti-HER2 antibody in tumors of mice. *Cancer Res.* **67,** 1138–1144 (2007).
34. Gao, X., Cui, Y., Levenson, R. M., Chung, L. W. & Nie, S. *In vivo* cancer targeting and imaging with semiconductor quantum dots. *Nature Biotechnol.* **22,** 969–976 (2004).
35. Shafer, M. W., Mangold, L., Partin, A. W. & Haab, B. B. Antibody array profiling reveals serum TSP-1 as a marker to distinguish benign from malignant prostatic disease. *Prostate* **67,** 255–267 (2007).

36. Fredriksson, S. *et al.* Protein detection using proximity-dependent DNA ligation assays. *Nature Biotechnol.* **20,** 473–477 (2002).

37. Kato, K., Ishimuro, T., Arima, Y., Hirata, I. & Iwata, H. High-throughput immunophenotyping by surface plasmon resonance imaging. *Anal. Chem.* **79,** 8616–8923 (2007).

38. Burg, T. P. *et al.* Weighing of biomolecules, single cells and single nanoparticles in fluid. *Nature* **446,** 1066–1069 (2007).

39. Kuhn, E. *et al.* Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and ^{13}C-labeled peptide standards. *Proteomics* **4,** 1175–1186 (2004).

40. Hoppe-Seyler, F. & Butz, K. Peptide aptamers: powerful new tools for molecular medicine. *J. Mol. Med.* **78,** 426-430 (2000).

41. Cerchia, L., Hamm, J., Libri, D., Tavitian, B. & de Franciscis, V. Nucleic acid aptamers in cancer medicine. *FEBS Lett.* **528,** 12–16 (2002).

42. Maser, R. S. *et al.* Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* **447,** 966–971 (2007).
    This paper describes the highly concordant nature of mouse cancer models and the human cancer genome.

43. Lu, H., Knutson, K. L., Gad, E. & Disis, M. L. The tumor antigen repertoire identified in tumor-bearing *Neu* transgenic mice predicts human tumor antigens. *Cancer Res.* **66,** 9754–9761 (2006).

44. Chen, R. *et al.* Pancreatic cancer proteome: the proteins that underlie invasion, metastasis, and immunologic escape. *Gastroenterology* **129,** 1187–1197 (2006).

45. Chen, R. *et al.* Quantitative proteomics analysis reveals that proteins differentially expressed in chronic pancreatitis are also frequently involved in pancreatic cancer. *Mol. Cell. Proteomics* **6,** 1331–1342 (2007)

46. Chen, R. *et al.* Comparison of pancreas juice proteins from cancer versus pancreatitis using quantitative proteomic analysis. *Pancreas* **34,** 70–79 (2007).

47. Lin, Y. *et al.* Proteins associated with disease and clinical course in pancreas cancer: a proteomic analysis of plasma in surgical patients. *J. Proteome Res.* **5,** 2169–2176 (2006).

48. Yu, K. H., Rustgi, A. K. & Blair, I. A. Characterization of proteins in human pancreatic cancer serum using differential gel electrophoresis and tandem mass spectrometry. *J. Proteome Res.* **4,** 1742–1751 (2005).

49. Wang, Y., Wu, S. L. & Hancock, W. S. Approaches to the study of N-linked glycoproteins in human plasma using lectin affinity chromatography and nano-HPLC coupled to electrospray linear ion trap–Fourier transform mass spectrometry (LTQ/FTMS). *Glycobiology* **6,** 514–523 (2006).

Imaging in the era of molecular oncology

Ralph Weissleder[1,2] & Mikael J. Pittet[1]

New technologies for imaging molecules, particularly optical technologies, are increasingly being used to understand the complexity, diversity and *in vivo* behaviour of cancers. 'Omic' approaches are providing comprehensive 'snapshots' of biological indicators, or biomarkers, of cancer, but imaging can take this information a step further, showing the activity of these markers *in vivo* and how their location changes over time. Advances in experimental and clinical imaging are likely to improve how cancer is understood at a systems level and, ultimately, should enable doctors not only to locate tumours but also to assess the activity of the biological processes within these tumours and to provide 'on the spot' treatment.

Imaging has become an indispensable tool in cancer research, clinical trials and medical practice. In the past three decades, there has been a huge increase in the number of imaging technologies and their applications (Fig. 1), but several challenges remain. At present, molecular imaging systems enable doctors to see where a tumour is located in the body. Ultimately, it is hoped that some of these systems will also help doctors to visualize the expression and activity of particular molecules, cells and biological processes that influence the behaviour of tumours and/or responsiveness to therapeutic drugs.

Perhaps the biggest growth area is fluorescence imaging, with various technologies being adapted for *in vivo* analysis. Indeed, because of developments in fluorescent imaging, researchers are now on the verge of being able to address some of the big questions in molecular oncology. How do the components of intracellular signalling pathways interact in real time? What are the kinetics and flux rates of such networks? What are the differences between networks in malignant cells and normal cells? Can we exploit these differences to make drugs that are less toxic and more efficacious? And what are the 'hubs' (the crucial signalling nodes) that will translate into the most efficient read-outs of how a cancer is progressing and whether a therapy is effective? In this article, we highlight recent advances in molecular imaging, with an emphasis on fluorescence imaging. We first provide an overview of the imaging technologies and then focus on specific applications in cancer biology and oncology. Cellular nanoimaging[1,2] and clinical imaging technologies[3–5] have been reviewed recently and so will not be covered in depth.

Overview of imaging technologies

Imaging systems can be grouped by the energy used to derive visual information (X-rays, positrons, photons or sound waves), the spatial resolution that is attained (macroscopic, mesoscopic or microscopic) or the type of information that is obtained (anatomical, physiological, cellular or molecular). Macroscopic imaging systems that provide anatomical and physiological information are now in widespread clinical and preclinical use: these systems include computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound. By contrast, systems that obtain molecular information are just emerging, and only some are in clinical and preclinical use: these systems include positron-emission tomography (PET), single-photon-emission CT (SPECT), fluorescence reflectance imaging, fluorescence-mediated tomography (FMT), fibre-optic microscopy, optical frequency-domain imaging,

bioluminescence imaging, laser-scanning confocal microscopy and multiphoton microscopy.

For imaging technologies to be adapted more widely and to be complementary to other types of molecular measurement, the read-outs need to meet certain criteria: they need to be quantitative, high resolution, longitudinal (that is, allow imaging over time), comprehensive, standardized, digital and sensitive to molecular perturbations in the system. In terms of quantification, imaging technologies can provide data that are absolute or relative. Absolute quantification is possible in techniques where signals are independent of position in the sample. CT, FMT, MRI and PET inherently provide quantifiable information. Relative quantification is obtained from image data sets whose signals are depth and sample-type dependent but can be validated by rigorous experimental design. Bioluminescence imaging, fluorescence reflectance imaging and multiphoton microscopy technologies belong in this category. Table 1 summarizes the spatial resolution, depth penetration, imaging time and cost of available systems.

For routine clinical practice and for testing the efficacy of drugs in clinical trials, CT, MRI, PET and SPECT are useful. Adaptations of these systems with much higher spatial resolutions have become available for use in experimental mouse models, allowing the development of new imaging probes for use in the clinic. By contrast, fluorescence reflectance imaging, FMT, fibre-optic microscopy and optical frequency-domain imaging are still mainly used experimentally, but they have clear potential for translation into the clinic. Because each technology has unique strengths and limitations, platforms that combine several technologies — such as PET–CT, FMT–CT, FMT–MRI and PET–MRI — are emerging, and these multimodal platforms have improved the reconstruction and visualization of data.

Complementary approaches to imaging include the use of portable *in vivo* flow cytometers and molecular 'nanolabs' to track circulating tumour cells (Box 1), as well as implantable, miniaturized fibre-optic multiphoton microscopy systems and implantable sensors for imaging molecular information in tumour environments. Other imaging technologies have recently been described, including thermal, electromagnetic and terahertz imaging, but their use *in vivo* or in oncology is not as established as the techniques described here.

Key developments in optical imaging

Photons travelling through tissue and interacting with tissue components form the basis of optical imaging techniques. In this section, we

[1]Center for Systems Biology, Massachusetts General Hospital, 185 Cambridge Street, CPZN 5206, Boston, Massachusetts 02114, USA. [2]Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Alpert 536, Boston, Massachusetts 02115, USA.

cover systems for the observation of molecular targets at the microscopic and macroscopic levels, as well as non-fluorescence-based optical imaging systems.

Fluorescence imaging

Fluorescence illumination and observation has been one of the most rapidly adapted imaging technologies, in both medicine and biological sciences. Fluorescence has spurred the development of sophisticated microscopic, mesoscopic and macroscopic imaging systems, as well as miniaturized fibre-optic approaches, imaging probes and genetic reporter systems. Fluorescence refers to the property of certain molecules

to absorb light at a particular wavelength and to emit light of a longer wavelength after a brief interval known as the fluorescence lifetime. The phenomenon of fluorescence was known by the middle of the nineteenth century and was then adapted to fluorescence microscopy in the early twentieth century. Medical applications of fluorescence imaging date back to 1924, when the autofluorescence of endogenous porphyrins was observed in tumours illuminated with ultraviolet light[6]. In 1942, red fluorescence by tumours was observed after intravenous administration of porphyrins[7], and the first use of fluorescein (which emits green light) to improve the detection of brain tumours was reported in 1948 (ref. 8). Britton Chance and others pioneered the reconstruction of
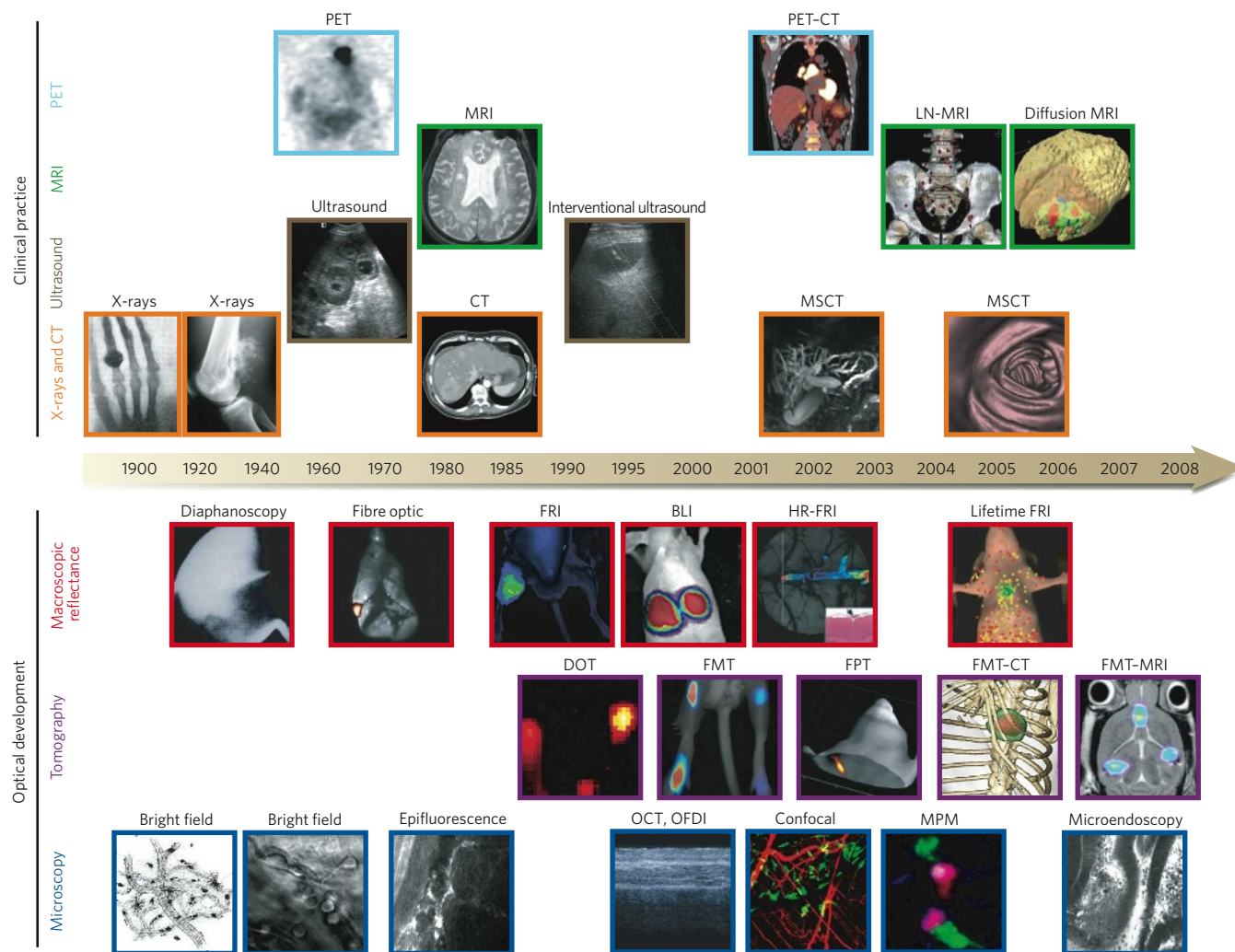
Figure 1 | Imaging technologies used in oncology. Many macroscopic imaging technologies (shown above the timeline) are in routine clinical use, and there have been huge advances in their capabilities to obtain anatomical and physiological information since the beginning of the twentieth century. Shown are some examples of bones (X-rays), soft tissue (ultrasound, MRI and CT rows), three-dimensional organs (CT and MRI rows) and physiological imaging (MRI and PET rows). Microscopic and other intravital optical techniques (shown below the timeline) have developed over the past decade and now allow studies of genetic, molecular and cellular events *in vivo*. Shown are surface-weighted, whole-mouse, two-dimensional techniques (macroscopic reflectance row); tomographic three-dimensional techniques, often in combination with other anatomical modalities (tomography row); and intravital microscopy techniques (microscopy row). The timeline is approximate and is not to scale. BLI, bioluminescence imaging; CT, computed tomography; DOT, diffuse optical tomography; FMT, fluorescence-mediated tomography; FPT, fluorescence protein tomography; FRI, fluorescence reflectance imaging; HR-FRI, high-resolution FRI; LN-MRI, lymphotropic nanoparticle-enhanced MRI; MPM, multiphoton microscopy; MRI, magnetic resonance imaging; MSCT, multislice CT;

OCT, optical coherence tomography; OFDI, optical frequency-domain imaging; PET, positron-emission tomography. (PET image reproduced, with permission, from ref. 4. Diffusion MRI image courtesy of B. Ross and A. Rehemtulla, Univ. Michigan Medical School, Ann Arbor. X-ray image (left) reproduced, with permission, from ref. 86. Diaphanoscopy image reproduced, with permission, from ref. 87. Fibre-optic image reproduced, with permission, from ref. 88. BLI image courtesy of K. Shah, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts. Lifetime FRI image courtesy of U. Mahmood and C. Salthouse, Massachusetts General Hospital. DOT image reproduced, with permission, from ref. 89. FPT image courtesy of G. Zacharakis and V. Ntziachristos, Massachusetts General Hospital. FMT–MRI image courtesy of J. Chen, Massachusetts General Hospital. Bright-field image (left) reproduced, with permission, from ref. 14. Bright-field image (right) courtesy of T. Mempel, Massachusetts General Hospital. Epifluorescence image courtesy of F. Jaffer, Massachusetts General Hospital. OCT, OFDI image reproduced, with permission, from ref. 22. MPM image reproduced, with permission, from ref. 51. Microendoscopy image reproduced, with permission, from ref. 17.)

Table 1 | Overview of imaging systems

Technique	Resolution*	Depth	Time†	Quantitative‡	Multi-channel	Imaging agents	Target	Cost*§	Main small-animal use	Clinical use
MRI	10–100 μm	No limit	Minutes to hours	Yes	No	Paramagnetic chelates, magnetic particles	Anatomical, physiological, molecular	$$$	Versatile imaging modality with high soft-tissue contrast	Yes
CT	50 μm	No limit	Minutes	Yes	No	Iodinated molecules	Anatomical, physiological	$$	Imaging lungs and bone	Yes
Ultrasound	50 μm	cm	Seconds to minutes	Yes	No	Microbubbles	Anatomical, physiological	$$	Vascular and interventional imaging||	Yes
PET	1–2 mm	No limit	Minutes to hours	Yes	No	^{18}F-, ^{64}Cu- or ^{11}C-labelled compounds	Physiological, molecular	$$$	Versatile imaging modality with many tracers	Yes
SPECT	1–2 mm	No limit	Minutes to hours	Yes	No	^{99m}Tc- or ^{111}In-labelled compounds	Physiological, molecular	$$	Imaging labelled antibodies, proteins and peptides	Yes
Fluorescence reflectance imaging	2–3 mm	<1 cm	Seconds to minutes	No	Yes	Photoproteins, fluorochromes	Physiological, molecular	$	Rapid screening of molecular events in surface-based disease	Yes
FMT	1 mm	<10 cm	Minutes to hours	Yes	Yes	Near-infrared fluorochromes	Physiological, molecular	$$	Quantitative imaging of fluorochrome reporters	In development
Bioluminescence imaging	Several mm	cm	Minutes	No	Yes	Luciferins	Molecular	$$	Gene expression, cell and bacterium tracking	No
Intravital microscopy¶	1 μm	<400–800 μm	Seconds to hours	No	Yes	Photoproteins, fluorochromes	Anatomical, physiological, molecular	$$$	All of the above at higher resolutions but limited depths and coverage	In development#

*For high-resolution, small-animal imaging systems. (Clinical imaging systems differ.) †Time for image acquisition. ‡Quantitative here means inherently quantitative. All approaches allow relative quantification. §Cost is based on purchase price of imaging systems in the United States: $, <US$100,000; $$, US$100,000–300,000; $$$, >US$300,000. ||Interventional means used for interventional procedures such as biopsies or injection of cells under ultrasound guidance. ¶Laser-scanning confocal or multiphoton microscopy. #For microendoscopy and skin imaging. (Table adapted, with permission, from ref. 85.)

light through tissue in the 1980s (ref. 9), on the basis of the observation that near-infrared photons (650–900 nm) travel through tissue much more efficiently than those in the visible range[10]. Concomitantly, fluorescence imaging with targeted molecular probes was developed[11–13]. Today, tomographic reconstruction of near-infrared photons, spectral unmixing, image fusion and multichannel imaging have become routine, and several preclinical and clinical imaging systems are commercially available.

Microscopic fluorescence imaging

Microscopic imaging has been used to study the activity of cells in many biological settings, including tumours, by a variety of illumination techniques. In 1839, Rudolph Wagner visualized leukocytes rolling in blood vessels within membranous translucent tissues by using bright-field transillumination[14]. With the advent of fluorescence microscopy, it then became possible to analyse multiple cell types simultaneously and in deep positions in solid tissues. Several imaging approaches based on fluorescence microscopy that were established for visualizing cells *in vitro* have recently been adapted for *in vivo* imaging: multiphoton microscopy, laser-scanning confocal microscopy, fibre-optic approaches and spectrally encoded endoscopy.

Multiphoton microscopy imaging systems achieve depth resolutions of up to 800 μm, yield three-dimensional information from light emitted by differentially labelled fluorescent objects (detected through multiple channels) and from second harmonic signals, and can provide hours of imaging at high spatial resolutions. Intravital (*in vivo*) multiphoton microscopy provides relative quantification of fluorescence signal (as described earlier), but it can be used to derive truly quantitative parameters of intravascular and interstitial cell migration (such as velocity, displacement, persistence time, and chemotactic and motility indices)[15]. In the frequently used animal models, cells are investigated in the popliteal and inguinal lymph nodes, the cranial bone marrow, tumour window chambers surgically implanted in the dorsum of the animal, and some exteriorized organs harbouring orthotopic cancers. Ancillary techniques such as second harmonic generation and fluorescence resonance energy transfer (FRET) are also increasingly being adopted for *in vivo* imaging, because they can add structural or

molecular detail. Techniques to quantify the FRET level include imaging after photobleaching or fluorescence lifetime.

Confocal microscopy set-ups have also become more widespread, because they are user-friendly and less costly than multiphoton microscopy. But cell death as a result of phototoxicity can be a limitation if observation times are lengthy, and optical penetration into tissues is lower.

Fibre-optic approaches have been adapted for intravital microscopy recently, potentially allowing imaging at orthotopic sites. Miniaturized systems are available for experimental work, and several other systems are being adapted for endoscopic clinical imaging[16,17].

Finally, spectrally encoded endoscopy is another new three-dimensional endoscopic technique that allows miniaturization of endoscopes, thus overcoming many of the limitations of the fibre-optic bundles of conventional endoscopes[18].

Macroscopic fluorescence imaging

Macroscopic fluorescence imaging systems rely on photographic principles to collect images in low light. There are two main types of imaging approach: fluorescence reflectance and tomographic fluorescence.

Fluorescence reflectance imaging systems consist of an excitation source, filters and a charge-coupled-device camera to obtain two-dimensional (planar) images. They are useful for imaging events in surface tumours (such as xenografts) and surgically exposed organs, and for intra-operative use. But they have a limited depth resolution beyond 3–5 mm and are not inherently quantitative.

Tomographic fluorescence systems[19] (FMT, FMT–CT, FMT–MRI and fluorescence protein tomography) reconstruct three-dimensional maps of fluorochromes on the basis of sophisticated algorithms, and these systems are quantitative. The early tomographic fluorescence systems used mice immersed in index-matching fluids to simplify the theoretical constraints associated with boundary modelling and to calculate light propagation. More recent 'free-space' systems eliminate the need for index-matching fluid. FMT can now also be combined with CT or MRI for improved photon reconstruction and image visualization[20]. Tomographic fluorescence systems have also been adapted to reconstruct visible fluorescence (for example, fluorescence protein tomography), thus enabling fluorescent proteins or genetically modified cells


Box 1 | Quantification of circulating tumour cells

The number of circulating tumour cells (CTCs) is a sensitive biomarker for tumour progression and metastasis[94]. Therefore, the quantification of CTCs is emerging as useful for diagnosing and 'staging' cancer, for assessing responses to treatment, and for evaluating whether there is residual disease. To be clinically useful, emerging technologies need to be highly sensitive (with detection limits around 1 cell per ml of blood) and highly specific for CTCs. Several such methods have recently been described and are in use experimentally.

Imaging CTCs non-invasively in microvessels of the skin could improve the sensitivity of detection by allowing large volumes of blood to be analysed. Confocal detection of CTCs with a dedicated *in vivo* flow cytometer was first demonstrated in 2004 (ref. 95), and portable systems have been described more recently[96]. In the figure, part **a** shows circulating green fluorescent protein (GFP)-positive multiple myeloma cells detected by *in vivo* flow cytometry in ear arterioles. Measurements were taken over 60 s at various intervals after tumour cells were injected (three separate days are shown). Each signal spike represents a single CTC. Therefore, the data show an increase in the numbers of CTCs over time, and this occurred together with tumour progression. Another technique that has been advocated is using multiphoton microscopy to image the peripheral vasculature after intravenous injection of tumour-cell-specific fluorescent ligands, such as fluorescent folates, which are thought to be internalized by tumour cells[97].

An alternative blood-screening method recently reported uses a highly sensitive microfluidic platform. The platform consists of an array of microposts that are made chemically functional with antibodies specific for epithelial cell-adhesion molecule (EpCAM) and therefore captures CTCs of epithelial origin[98] (not shown).

Finally, a chip-based diagnostic MRI (DMR) platform with multiple
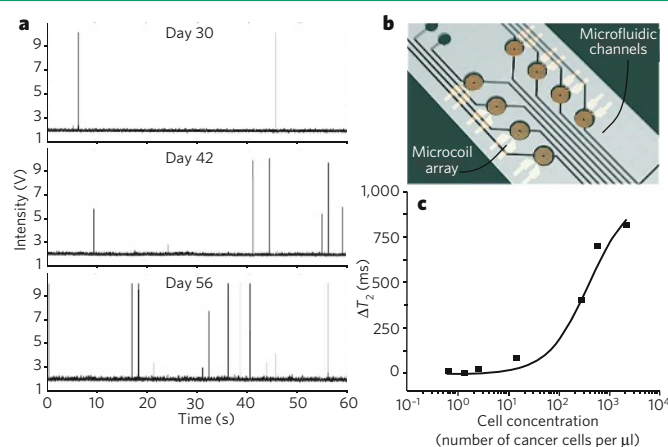
channels allows rapid and quantitative detection of biological targets[99], as shown in parts **b** and **c** of the figure. Using functionalized magnetic nanoparticles as proximity sensors to amplify molecular interactions[100], the DMR system can carry out highly sensitive and selective measurements on small volumes of unprocessed biological samples, including the profiling of circulating cells and the multiplexed identification of cancer biomarkers[99]. Part **c** of the figure shows detectable DMR changes (change in the time of magnetic relaxation, ΔT_2) of whole blood as a function of added cancer cells. (Panel **a** courtesy of C. Lin, Massachusetts General Hospital, Boston, Massachusetts, and A. Kung and I. Ghobrial, Dana-Farber Cancer Institute, Boston, Massachusetts. Panel **b** adapted, with permission, from ref. 99.)

---

to be tracked irrespective of cell division *in vivo*[21]. New fluorescence protein tomography systems can image organisms (such as fruitflies and nematodes) at mesoscopic resolution.

Non-fluorescence-based optical imaging

Several optical imaging systems that are not based on fluorescence have also been introduced: bioluminescence imaging, optical coherence tomography, photoacoustic microscopy and tissue spectroscopy.

Bioluminescence imaging has emerged as a useful technique for imaging small experimental animals. The imaging signal depends on the expression levels of a luciferase enzyme, the presence of exogenously administered enzyme substrate (a luciferin), ATP, oxygen and depth. Numerous luciferase–luciferin pairs have been harnessed for *in vivo* imaging. Particularly useful pairs are firefly (*Photinus pyralis*) luciferase–luciferin, because of the high wavelength of emitted light and quantum yields, and *Renilla reniformis* luciferase–coelenterazine and *Gaussia princeps* luciferase–coelenterazine, because of their flash kinetics and ability to generate photons outside cells. Typical doses of luciferin are in large excess (>100 mg per kg body weight administered intraperitoneally) and are injected immediately before data acquisition. In contrast to fluorescence techniques, there is no inherent background noise, which makes this technique highly sensitive. However, at present, this method does not allow absolute quantification of target signal; instead, its primary uses are in binary mode (that is, yes or no for luciferase expression) or as an imaging tool to follow the same animal over time in identical conditions, including positioning of the body.

Optical coherence tomography is based on light scattering and can be used to image microscopic structures *in vivo* (at a resolution of 2–15 µm and to a depth of 1–3 mm). Studies in humans have shown that this method identifies dysplasia and colonic adenomas. Until now, however, it has been too slow to provide comprehensive microscopic imaging and has therefore been used as a point-sampling technique with a field of view comparable to that in a conventional biopsy. Optical frequency-domain imaging leverages the high sensitivity of optical coherence tomography but is several orders of magnitude faster[22] and can thus be used to screen larger organ samples for the presence of malignancies.

Photoacoustic microscopy uses short laser pulses to irradiate tissue and temporarily raise its temperature (by millikelvins). Thermo-elastic expansion then causes the emission of photoacoustic waves that can be measured by wide-band ultrasonic transducers, offering improved depth resolution in the 3–20 mm range[23]. The technique combines ultrasonic-scale spatial resolution with high sensitivity to tissue light absorption and can yield information on physiology or on exogenously administered light absorbers. Newer generations of imaging systems provide macroscopic tomographic capabilities (a process known as photoacoustic tomography) or use radio-frequency pulses to generate acoustic waves (a process known as thermal acoustic tomography).

Finally, tissue spectroscopy detects relative changes in the way in which light interacts with tissue and has been used extensively to improve early detection of gastrointestinal malignancies[24]. The vibrational spectra of biological specimens can be used to identify the biochemical constituents of tissue, but there are practical limitations of this technique for *in vivo* use, owing to its relatively low sensitivity and limited spatial and temporal resolutions. Much stronger vibrational signals can be obtained with coherent anti-Stokes Raman scattering (CARS), a nonlinear Raman technique[25]. CARS microscopy is useful for mapping lipid compartments, protein clusters and water distribution. Other optical contrast mechanisms such as absorption, polarization, coherence, interference and elastic (or Mie) scattering can also be exploited to generate images, some of which might be particularly useful for oncological applications.

Labelling

With few exceptions, cells, whether they are normal cells or cancer cells, cannot easily be distinguished from each other by *in vivo* imaging. Consequently, the molecules and cells of interest need to be labelled to become visible. Labelling techniques can be broadly divided into three categories: genetic reporters, injectable imaging agents and exogenous cell trackers.

Genetic reporters

Genetic reporter strategies have largely been limited to mouse models of cancer, but their use is widespread in basic biological sciences[26,27]. The

types of protein expressed by the reporter genes used in optical imaging are fluorescent proteins (green, far-red and photoswitchable fluorescent proteins), bioluminescent proteins (luciferases from several organisms) or other proteins (such as herpes simplex virus 1 thymidine kinase, transferrin receptor, somatostatin receptor type 2, dopamine receptor D2 and human Na^+/I^- symporter). The optical reporter genes are cloned into the promoter or enhancer region of a gene of interest or engineered to be expressed as a fusion with the protein of interest, allowing longitudinal studies of biological processes. Such reporters are particularly valuable in genetically engineered mouse models in which the activation of oncogenes and/or tumour-suppressor genes is under temporal and tissue-specific control, because the reporters enable not only oncogenic transformation to be studied but also drug action, toxicity and resistance[28,29].

Three genetic reporter strategies are available to study protein–protein interactions *in vivo*: modified mammalian two-hybrid systems[30]; bioluminescence resonance energy transfer (BRET)[31] and FRET[32]; and split reporter[33]. Emerging two-hybrid systems are fairly simple and rapid, but they still need to be optimized. The reaction occurs in the nucleus, so the system can mainly be used to interrogate interactions between nuclear proteins or proteins that are modified to translocate to the nucleus. BRET, unlike FRET, usually does not cause photodamage to the cells and does not require excitation of the donor with an external light source, so it has less background noise, which might ultimately translate into being able to detect protein–protein interactions at lower concentrations. Split-reporter strategies, in which a reporter gene is split into components that become active when reconstituted, generate specific signals and can be used to image interacting proteins anywhere in the cell, but they need to be adapted for reversible interactions. These three genetic reporter strategies also allow monitoring of the concentrations of ions, sugars, lipids and other active molecules *in vivo*, so they could further broaden the application of intravital microscopy. Using a combination of the above genetic-reporter strategies for labelling should increase our knowledge of entire biological pathways and accelerate a systems-wide understanding of biological complexity.

Injectable imaging agents

Injectable imaging agents that are specific for molecular targets have the advantage of being usable in both experimental animals and humans. But, at present, such agents exist for <5% of the molecular targets of interest in cancer. This is largely because *in vivo* imaging agents face more stringent design criteria than *in vitro* reagents. The basic underlying difficulty is to design agents with high target-to-background ratios. The ideal agent should contribute minimally to the background signal (for example, there should be minimal non-specific tissue extravasation, internalization by macrophages, or renal or hepatic elimination, which obscures adjacent organs) but yield high local concentrations at the intended sites. For these reasons, *in vivo* agents are often designed using chemical or biological amplification strategies. Chemical amplification strategies include using multiple copies of a label in a single molecule (multivalency), pairing of deactivators and fluorochromes (quenching), using covalent target binders (active-site binding), inducing magnetic changes on target interaction (relaxivity changes and magnetic switches), releasing caged compounds by photolysis (uncaging) and using distance-dependent interaction between excited states of dyes (FRET). Biological amplification strategies include intracellular sequestration of the imaging agent (cellular trapping), generation of signal by endogenous enzymes in target cells (enzymatic conversion), and a two-step activation process in which molecular agents are tagged first by a multivalent agent and then by a fluorescent conjugate (pre-targeting).

Some generic types of imaging probe are summarized in Fig. 2. Of particular interest in the past few years have been high-throughput methods such as phage display[34]; systematic evolution of ligands by exponential enrichment (SELEX); fragment-based, template-based and nanoparticle-based[35] synthesis for library design; and fast *in vivo* screening techniques[36]. Spectrally resolved imaging agents should foster

systems approaches that allow the evaluation of multiple cellular and/or molecular activities *in situ*.

Exogenous cell trackers

Exogenous cell trackers are being used increasingly to study adoptive cell-transfer therapies. Recent translational approaches for cell labelling include an [^{111}In]oxine tracer[37] (which is approved by the US Food and Drug Administration), derivatized magnetic nanoparticles[38–40], and novel amine-reactive esters of benzindolium-derived fluorochromes for macroscopic and microscopic imaging[41].

Applications in oncology

In this section, we review the roles in tumorigenesis that have been ascribed to various cells and molecules on the basis of *in vivo* studies. We also highlight some of the main applications of, and recent advances in, tumour imaging. Several other review articles have discussed the use of established clinical imaging technologies[3–5].

Tumours and the host response

Stimuli from tumours affect components of the extracellular matrix (ECM) (such as collagen, fibronectin and glycoproteins) and cells present in the environment of the tumour (such as T cells, macrophages, dendritic cells, mast cells, neutrophils, fibroblasts and endothelial cells). These ECM components and cells can have complex regulatory functions mediated by proteolytic enzymes, cytokines, growth factors and angiogenesis-promoting factors. Recent advances in intravital imaging have made it possible to quantify and catalogue the behaviour and functions of some of the cellular and molecular components of the immune system that control tumour growth (Fig. 3). Importantly, intravital microscopy studies indicate that the behaviour of immune cells in tissues is dictated by local factors in the organ that often cannot be reproduced *in vitro* (at least for the foreseeable future).

Antitumour cytotoxic T cells (CTLs) that recognize short antigenic peptides presented by tumour cells can kill their targets in an antigen-dependent and contact-dependent manner by releasing preformed cytotoxic granules or by generating potent inflammatory responses through the secretion of cytokines[42,43]. Other immune cells suppress the function of these CTLs, thus promoting tumour growth: these include $FOXP3^+$ regulatory T (T_{reg}) cells[44,45] and tumour-associated macrophages (TAMs)[46,47].

The immunological functions mediated by effector and suppressor cells have mostly been characterized *in vitro*. *In vivo* studies at cellular resolution have been carried out only recently. A recent finding from two intravital multiphoton microscopy studies is that CTLs are recruited from blood vessels in the tumour periphery and then infiltrate towards the centre of the tumour[48,49]. Sustained locomotion of tumour-infiltrating CTLs requires the presence of cognate antigen (that is, the particular antigen recognized by each CTL); thus, only CTLs specific for tumour antigens can infiltrate tumours deeply. In conditions in which CTLs efficiently reject tumours, the CTLs migrate randomly throughout the tumour environment at high instantaneous velocities and pause to form antigen-dependent interactions with tumour cells. The dynamics of CTL–tumour cell interactions are diverse: some tumour-specific CTLs engage in long-lasting interactions, whereas others establish sequential, short-term contacts. The reasons for these different behaviours are unknown, but during this interaction time, CTLs could be delivering their cytotoxic granules and/or integrating signals for the production of pro-inflammatory cytokines. The progeny of CTLs that have come into contact with antigen typically include cells with distinct effector functions[50], which might have diverse modes of interaction with tumour cells *in vivo*. When CTLs detach from their tumour-cell targets, they become highly motile again, presumably in search of intact tumour cells.

The presence of T_{reg} cells can compromise killing by CTLs and prevent efficient tumour rejection. Indeed, another intravital multiphoton microscopy study has shown that CTLs have a constrained window of opportunity to induce target-cell death before the cell–cell conjugates dissociate (typically about 15 min), and T_{reg} cells suppress the function of tumour-specific CTLs by delaying granule exocytosis beyond this

time frame (typically >30 min)[51]. In the presence of these T_{reg} cells, the CTLs behave otherwise 'normally' *in situ*: they proliferate extensively, are highly motile, contain cytotoxic granules, and form antigen-dependent conjugates with target cells. Interestingly, T_{reg}-cell-mediated suppression is reversible, because *in vivo* depletion of T_{reg} cells quickly restores CTL function[51]. These findings have clinical significance because patients with cancer also accumulate T_{reg} cells in tumour stroma[44], where CTLs fail to exert a cytotoxic function[52]. Intravital imaging studies in mouse models indicate that T_{reg}-cell-mediated suppression requires cytokines such as transforming growth factor-β but not prolonged interactions with CTLs[51]. Because T_{reg} cells are relatively poor producers of transforming growth factor-β, they might function by mobilizing cellular intermediates, such as TAMs. Indeed, CTLs in the tumour stroma form long-lasting interactions with TAMs[49], which might supply a variety of immunosuppressive signals. Clearly, more intravital microscopy studies are needed to identify precisely the protumorigenic and antitumorigenic activities of the various cells in the tumour stroma.

Tumour invasion and metastasis

The motility of tumour cells is an essential feature for intravasation at primary sites and extravasation at distant sites to form metastases. For cells to migrate, they need to adhere to selected substrates, to reorganize their cytoskeleton, and to acquire the spatial asymmetry that enables them to turn intracellularly generated forces into locomotion. A core set of proteins, including those of the cofilin pathway, coordinates actin polymerization and is essential for the formation of cellular protrusions

and crawling. Multiphoton microscopy imaging and gene-expression profiling in a mouse model of breast cancer have shown that the expression of genes in the cofilin pathway is altered in tumour cells and that these cells show dysregulated chemotaxis towards environmental stimuli, such as epidermal growth factor (EGF) and the chemokine CXCL12 (ref. 53).

Imaging studies have also confirmed that tumour-cell migration and invasion depend on stromal cells such as TAMs and carcinoma-associated fibroblasts (CAFs)[54–56] (Fig. 3). By visualizing differentially labelled tumour cells and TAMs *in situ*, it has been shown that motile tumour cells preferentially migrate towards perivascular TAMs that reside deep within breast tumours. The tumour cells move efficiently along linear paths in association with proteoglycans that are present on ECM fibres, which might be guidance cues for movement towards blood vessels.

Tumour cells eventually exit into the blood vessels in proximity to TAMs; the efficiency of this egress depends on the number of perivascular TAMs but not on the number of blood vessels. It is possible that the EGF receptor (EGFR) at the surface of tumour cells responds directly to chemoattractants such as EGF that are secreted by perivascular TAMs. Indeed, inhibition of the EGFR in this model markedly decreases the number of tumour cells exiting into the blood[54]. Tumour cells can also secrete colony-stimulating factor 1 (CSF1; also known as M-CSF), which stimulates TAMs to secrete EGF, thereby further promoting tumour progression (through promoting intravasation, which can lead to metastasis). Together with TAMs, CAFs are present in large numbers in the stromal compartment of most invasive human breast cancers. CAFs secrete more CXCL12 than do fibroblasts in non-cancerous regions. The increased

Small ligands
Metabolic substrates, labelled drugs and peptides

Active-site binders
Bind to proteases, protein kinases and other enzymes

Site-specific protein tags
FlAsH tag, His tag and IQ-tag

Environmentally sensitive probes
Sense major histocompatibility complex molecules, pH and proteases

Supramolecular structures
Prosense and polymers

Engineered proteins
Antibody fragments, diabodies and triabodies

Inorganic nanoparticles
Magnetic nanoparticles, sparks and quantum dots

Bionanoparticles
Magnetic phage and fluorescent phage

Figure 2 | High-affinity imaging agents with appropriate pharmacokinetics are essential for imaging at the molecular level. Different strategies have been pursued to develop agents for molecular imaging, and the various types of agent available and examples of these are indicated. Small molecules are shown in the top row, and macromolecular agents and nanotechnology-derived agents are shown in the bottom row. Four main types of small molecule are used. Small ligands refers to imageable small molecules (for example, ^{18}F-labelled drugs and fluorescent peptides), whereas active-site binders (green) attach to specific protein pockets in enzymes (blue) either covalently or non-covalently. Site-specific protein tags (pink) achieve a similar function but at sites of interest in engineered proteins (white). Environmentally sensitive probes (for example, 4-*N*,*N*-dimethylamino-phthalimidoalanine (4-DAPA); yellow) change their physicochemical

properties on interaction with the target (in this case, Tyr and T90β). Two main types of macromolecule are used. Supramolecular structures are synthetic agents that have been useful as enzyme substrates for the imaging of cathepsins and proteases or for delineating new microvasculature: shown is part of a poly-L-lysine backbone (blue) derivatized with protease-cleavable side chains (red) and polyethylene glycol (grey). Engineered proteins (with optimized pharmacokinetics) refers to other macromolecules that have been used for some targets. Finally, a host of nanomaterials, including inorganic nanoparticles (grey) and bionanoparticles, that can be used for imaging phagocytic cells or cell-surface targets (green) is emerging. (Images adapted, with permission, from ref. 90 (active-site binders), ref. 91 (site-specific protein tags), ref. 92 (environmentally sensitive probes), Mimetibody.com (engineered proteins) and ref. 93 (inorganic nanoparticles).)

585

amount of CXCL12 has several effects: it promotes carcinoma cell growth by direct paracrine stimulation of the chemokine receptor CXCR4 on tumour cells; it recruits endothelial progenitor cells from the blood vessels, giving rise to highly vascularized tumours; and it supports the metastasis of tumour cells towards lymphatic vessels[55,57]. Other cells present in the tumour stroma, such as mast cells[58], neutrophils[59] and mesenchymal stem cells[60], can also promote cancer invasion and metastasis, but their activity *in situ* has not been investigated.

Matrix remodelling and angiogenesis

TAMs and tumour cells overproduce matrix metalloproteinases, plasminogen activators, cathepsins and other hydrolases. These enzymes degrade structural components of the ECM and thereby facilitate migration of the tumour cells, as well as that of endothelial cells and immune cells. The enzymes also change the nature of the interaction between cells and the ECM, promote cell growth, and promote the 'angiogenic switch' by liberating various ECM-bound cytokines. Angiogenesis — the formation of new blood vessels — provides oxygen and nutrients to tumours and promotes invasion and metastasis. There are several agents in development to image the key cells and effector molecules involved in these processes. Functional nanomaterials specific for macrophage subtypes have been reported[35], and several imaging agents also allow the visualization of proteases involved in ECM remodelling and the measurement of angiogenesis and associated parameters.
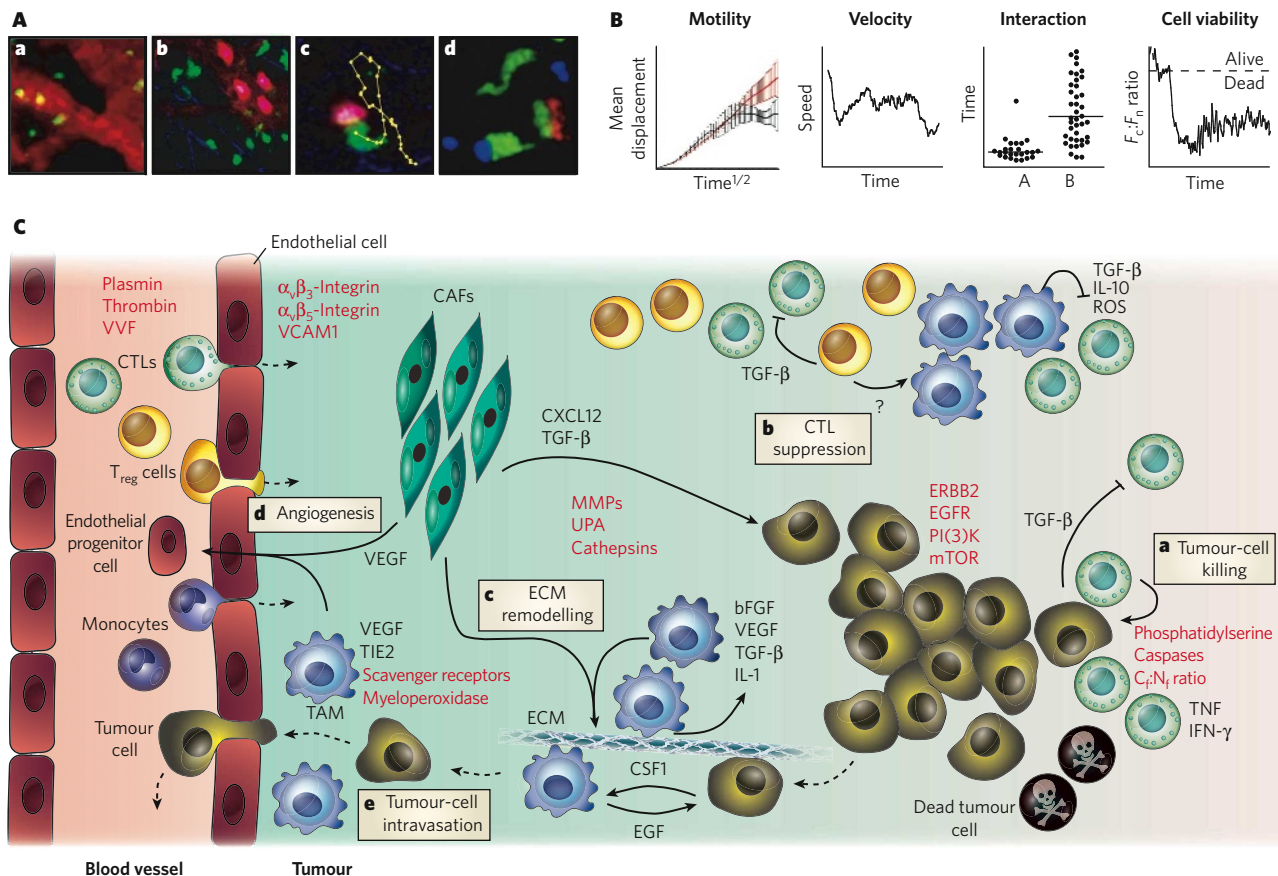
Figure 3 | Behaviour of tumour cells and cells in the tumour stroma *in situ*, as elucidated by *in vivo* imaging. Intravital microscopy can be used to derive quantitative parameters on intravascular and interstitial cell migration, interaction and viability in the tumour environment and the lymphatic organs. **A,** *In vivo* microscopic recordings in anaesthetized mice can be used to obtain data at cellular resolution. Some of the types of process that can be observed are illustrated. **a,** Leukocytes (green) trafficking in blood microvessels (red). **b,** B cells (pink) and T cells (green) inside and outside high-endothelial venules (diffuse red), respectively, in a tumour-draining lymph node. Collagen fibres are also visualized (blue). **c,** A cytotoxic T cell (CTL) (green) recognizing and killing a target cell (pink). Yellow dots track the position of the centre of the CTL at 15 s intervals. Collagen fibres are also visualized (blue). **d,** Simultaneous tracking of multiple cell types: a CTL (green) is firmly bound to an antigen-presenting cell (blue), and another CTL is transiently interacting with a FOXP3+ regulatory T (T_{reg}) cell (red). (Panels **b**–**d** reproduced, with permission, from ref. 51.) **B,** A typical analysis of cellular behaviour from data retrieved from time-lapse intravital imaging can assess four properties: the motility of cell populations (based on mean displacement over the square root of time); the instantaneous velocity of a single cell over time; the average cell–cell interaction time (in this case, A shows CTL–T_{reg}-cell interactions, and B shows CTL–target cell); and the viability of a single cell over time (as measured by the ratio of cytoplasmic fluorescence to nuclear fluorescence, F_c:F_n)[15,51]. **C,** A model for the *in vivo* role of tumour stromal cells in tumour growth and metastasis, based on intravital imaging studies, is shown. Immune cells such as monocytes (which differentiate into tumour-associated macrophages, TAMs), CTLs and T_{reg} cells are recruited into the tumour stroma from the blood vessels (dashed lines indicate direction) and can be visualized by intravital microscopy. Imaging studies can be used to investigate how these and other cells (such as carcinoma-associated fibroblasts, CAFs) interact with each other or with tumour cells and how they participate in tumorigenesis. In particular, *in vivo* studies have examined how CTLs recognize tumour-associated antigens at the surface of tumour cells[41,48,49] and exert cytotoxic activity[51] (**a**), how T_{reg} cells suppress the function of CTLs[51] (**b**), how TAMs and CAFs remodel the extracellular matrix (ECM) and promote invasion (**c**) and angiogenesis[54,56] (**d**), and how TAMs also participate in tumour-cell intravasation[54] (**e**). Macroscopic imaging techniques are complementary and allow bulk cell migration to be measured. Objects that can be imaged *in vivo* include the cell types described here, as well as molecular targets (or parameters) for which specific probes have been designed (indicated in red). bFGF, basic fibroblast growth factor; CSF1, colony-stimulating factor 1 (also known as M-CSF); CXCL12, CXC-chemokine ligand 12; EGF, epidermal growth factor; EGFR, EGF receptor; IFN-γ, interferon-γ; IL, interleukin; MMP, matrix metalloproteinase; mTOR, mammalian target of rapamycin; PI(3)K, phosphatidylinositol-3-OH kinase; ROS, reactive oxygen species; TGF-β, transforming growth factor-β; TIE2, endothelial tyrosine kinase (also known as TEK); TNF, tumour-necrosis factor; UPA, urokinase-type plasminogen activator; VCAM1, vascular cell-adhesion molecule 1; VEGF, vascular endothelial growth factor; VVF, vascular volume fraction.

Three-dimensional molecular sensing of protease activity *in vivo,* for example, is now feasible by combining activatable imaging probes based on biologically compatible near-infrared fluorochromes[61] and FMT for positional accuracy and detection in deep tissues[62]. These imaging probes can accumulate in tumours, because the probes have long half-lives in the blood and because the new blood vessels in the tumour are leaky; the proteases in the tumour stroma then cleave the probe, which consists of a labelled protein, thereby releasing the previously quenched fluorochromes locally. Using a combination of FMT and CT further improves signal localization and shows tumour burden; the procedure can be repeated serially (over time) to follow tumour progression[20]. The role of specific proteases in ECM degradation or other activities has not been investigated *in vivo* at subcellular resolution, although the protease sensors mentioned above will probably be considered for future multiphoton microscopy investigations.

Longitudinal intravital microscopy studies have also uncovered unique features of tumour blood vessels and angiogenesis[63,64]. For example, blood flow in tumours is highly variable, can reverse direction over time and does not deliver nutrients effectively. In many cancers, tumour blood vessels have defective endothelial barriers with relatively large pores, and pericytes, which normally provide support to endothelial cells, are often absent. These abnormalities, together with pressure from proliferating cells, contribute to high interstitial pressure and hypoxia in tumours. In addition, intravital microscopy studies[65–67] and tomographic fluorescence mapping[68] are being used to compare the efficacy of angiogenesis inhibitors and to investigate the efficacy of therapeutic drug delivery. Interestingly, studies of tumour blood-vessel responses have shown that certain antiangiogenic agents transiently normalize the structure and function of tumour vasculature, instead of 'pruning' the vessels as was expected[69]. This finding indicates that, at least in mouse models, there is a window of opportunity for improving drug delivery and for increasing sensitivity to radiation treatment.

Cell death

Real-time imaging of cell death would be a coveted application with which to assess the efficacy of cytotoxic drugs and, potentially, to monitor the toxicity of these drugs in normal tissues. Successful new-generation drugs are likely to be designed to exploit the differences (rather than the similarities) in activation of the apoptotic machinery between malignant and normal cells. The ability to image such effects *in vivo* could thus have far-reaching implications for assessing therapeutic efficacy and toxicity.

Bioluminescence imaging has been used to evaluate tumour growth and regression (for example, in response to therapy) in experimental animals with tumours engineered to express a luciferase[70], although this technique allows only relative quantification (as mentioned earlier). Strategies to image cell death directly include the use of recombinant photoprotein reporters activatable by caspase-3 cleavage[71,72], and fluorescent[73] or radiolabelled[74] ligands, such as annexin V, with high affinity for apoptotic cells. In fact, pharmaceutical companies are increasingly using whole-animal optical imaging systems, with tumour size or molecule-specific imaging signal as the primary read-outs for drug efficacy. Complementing the macroscopic information provided by bioluminescence imaging, recent approaches to image apoptosis at the single-cell level *in vivo* include labelling components in the cytoplasm and the nucleus (DNA) with different fluorochromes and then monitoring the cytoplasm:nucleus fluorescence ratio (F_c:F_n) over time; dead cells typically show a decrease in this ratio, owing to DNA fragmentation and loss of plasma-membrane integrity[51]. An alternative strategy is to visualize cellular morphology, because condensation of the cell body and blebbing of the plasma membrane are characteristic of cells entering apoptosis. These approaches have been used successfully to elucidate the *in vivo* dynamics of CTL-mediated killing and the mechanisms used by T_{reg} cells to blunt the cytotoxic functions of CTLs[49,51]. In the clinic, analysing glucose metabolism by using PET (after administering the probe [18F]fluorodeoxyglucose, which is preferentially taken up by metabolically active cells) is progressively being used as a surrogate for treatment efficacy[75,76].
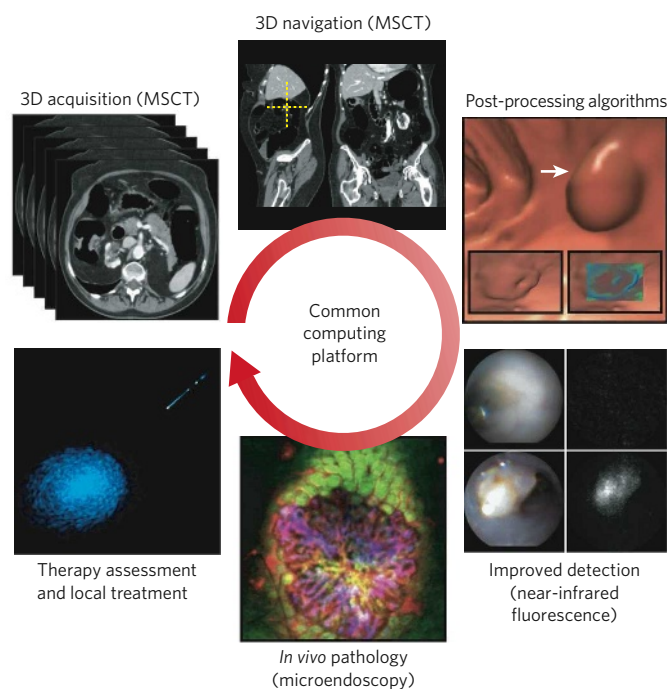
Figure 4 | Clinical imaging. With a common computing platform, data obtained by different imaging techniques can be seamlessly assembled (and fused) for the screening, detection, characterization (*in vivo* pathology) and real-time treatment of early-stage cancers. In the example shown, a multislice CT image of the abdomen is acquired, and the slices are assembled in three dimensions. A suspicious colonic lesion is identified (at the intersection of the dashed lines) and reconstructed in three dimensions to reveal a polyp-like growth (arrow). Post-processing algorithms that explore differences in tissue and air–barium attenuation (insets) can be used to suggest, but not confirm, the presence of a malignancy. Near-infrared endoscopy using protease beacons then allows the detection of small and otherwise difficult-to-identify lesions. The bright signal in the near-infrared fluorescence channel indicates large amounts of proteases in a tumour (lower right) but not in a normal colon (upper right). White-light views of the same images are shown on the left. Microendoscopy shows a malignant signature *in vivo*, which can be used to make treatment decisions on the spot in real time. The microscopy image shows a polyp-like lesion, which contains mucosal cells (green), proteases (red) and new microvasculature (blue). Real-time local treatment can be achieved by lesion removal or light-activated therapy. 3D, three-dimensional.

Clinical detection of epithelial neoplasias

New fibre-optic imaging technologies are being explored to detect early-stage epithelial cancers in the gastrointestinal tract[77] (Fig. 4), the lungs[78] and other easily accessible surfaces. These developments are important for several reasons. First, traditional endoscopy can fail to detect up to 20% of lesions during a procedure. Second, treatments for early-stage cancers are often curative, in contrast to later-stage treatments. Third, advanced imaging technologies can assist therapeutic endoscopic procedures (including laser-based surgery, cryotherapy, electrocautery and stenting). Last, emerging newer procedures could be used to differentiate premalignant or malignant lesions from inflammatory lesions in individuals who are at high risk of developing cancer.

Two clinical advances are particularly noteworthy given their high sensitivity and specificity: near-infrared fluorescence endoscopy, and fibre-optic confocal laser-scanning microendoscopy. Near-infrared fluorescence endoscopy, in combination with new imaging agents, is an emerging technology that has strong potential for increasing sensitivity[79]. The agents that are closest to clinical trials are fluorescent substrates that are activated by tumour-associated cathepsins[20,80]. Second, fibre-optic confocal laser-scanning microendoscopy has emerged as a tool for identifying cellular and subcellular structures on the basis of reflected light, autofluorescence or exogenous imaging agents. This endoscopic

technique allows diagnosis of colorectal cancer, helps to target intervention to relevant areas, and decreases the number of biopsies required for cancer surveillance[17]. It is expected that this combination of a macroscopic strategy and a microscopic strategy — that is, using near-infrared fluorescence endoscopy to identify suspicious lesions and then fibreoptic multichannel microendoscopy to analyse them microscopically — will ultimately allow 'virtual histology'.

These approaches herald the beginning of a new era that will give researchers and doctors a unique look at cellular structures and functions at and below epithelial surfaces. There is a clear need and opportunity for developing sets of biologically compatible, diagnostic near-infrared fluorescence imaging agents that are specific for different cell types and tissue structures and that enable multichannel measurements to be made, similar to the reagents available for *in vitro* fluorescence microscopy[34]. Finally, the capability of multichannel macroscopic and microscopic systems will allow immediate treatment decisions. Several agents that have combined diagnostic and therapeutic ('theranostic') properties are in development[81].

Intra-operative imaging

Fixed-geometry and hand-held fluorescence imaging systems should improve oncological surgical procedures. There are several emerging applications: defining tumour margins and identifying small metastases to improve the accuracy of surgical removal, mapping sentinel lymph nodes, and defining anatomy during surgical intervention. Most systems are based on reflectance imaging at video rate, combined with the use of near-infrared fluorochromes. Tumour margins have been visualized by using several fluorescence strategies, including targeting cancer cells[82], and targeting the proteolytic or phagocytic activities of tumour cells or host cells in the periphery[79,83]. Sentinel lymph-node mapping has been improved through the use of near-infrared fluorochrome-labelled albumin, fluorescent nanoparticles and red-shifted quantum dots[84].

Challenges for the future

In the past decade, enormous strides have been made in the imaging sciences, and many new technologies, agents and reporters have been applied to oncology research, clinical trials and patient care. What has facilitated these advances, and how do we ensure that they will continue in the future? Advances have been fuelled by the convergence of biology, chemistry, physics and engineering in interdisciplinary centres across the world. Miniaturization, nanotechnology and computation have also been driving forces behind innovation. Commercial availability of new imaging systems and imaging probes has also contributed to the adoption of these techniques. One area that has lagged behind, however, is the creation of information technology infrastructures that serve macroscopic and microscopic imaging communities alike. Although the Digital Imaging and Communications in Medicine standards are common in medical imaging and are the backbone of picture archiving and communication systems and electronic patient records, such standards have not commonly been adopted by the *in vivo* microscopy community. There are additional opportunities in image archival, distribution, analysis, visualization and creation of standardized databases to compare and analyse data sets.

Imaging is one of the few technologies that can generate longitudinal data sets in intact host environments. As the research community starts to adopt a more systems-wide analysis, it is clear that the data will need to fulfil certain criteria: they need to be truly quantitative (that is, absolute not only relative), comprehensive and able to be integrated with other molecular data sets. Quantification and integration are some of the more immediate needs.

The translation of imaging agents and technologies into the clinic has been much slower than was initially hoped. The reasons for this are complex and include considerable regulatory hurdles, market forces, lower profit margins for imaging than for therapeutic drugs, and the lack of reimbursement strategies for newer imaging agents. Despite these hurdles, several new technologies are poised to enter clinical trials. With the deeper understanding of the molecular basis of disease

that has been gained by imaging studies — and the recent advances in detection technology, imaging systems, reconstruction algorithms and visualization tools — the transformative promise of imaging is likely to be fulfilled soon. Applying the new molecular imaging tools to humans will make a fundamental improvement in how cancer is understood *in vivo* and should allow earlier detection, stratification of patients for treatment, and objective evaluation of new therapies in a given patient. The outcome will be considerably better management and care of those with cancer. ∎

1. Soon, L., Braet, F. & Condeelis, J. Moving in the right direction — nanoimaging in cancer cell motility and metastasis. *Microsc. Res. Tech.* **70,** 252–257 (2007).
2. Deisseroth, K. *et al.* Next-generation optical technologies for illuminating genetically targeted brain circuits. *J. Neurosci.* **26,** 10380–10386 (2006).
3. Brindle, K. M. New approaches for imaging tumour responses to treatment. *Nature Rev. Cancer* **8,** 94–107 (2008).
4. Quon, A. & Gambhir, S. S. FDG-PET and beyond: molecular breast cancer imaging. *J. Clin. Oncol.* **23,** 1664–1673 (2005).
5. Neves, A. A. & Brindle, K. M. Assessing responses to cancer therapy using molecular imaging. *Biochim. Biophys. Acta* **1766,** 242–261 (2006).
6. Policard, A. Étude sur les aspects offerts par des tumeurs expérimentales examinées à la lumière de Wood. *C. R. Séances Soc. Biol. Fil.* **91,** 1423–1424 (1924).
7. Auler, H. & Banzer, G. Untersuchungen ueber die Rolle der Porphyrine bei geschwulstranken Menschen und Tieren. *Z. Krebsforsch.* **53,** 65–68 (1942).
8. Moore, G. E., Peyton, W. T. & French, L. A. The clinical use of fluorescein in neurosurgery. The localization of brain tumors. *J. Neurosurg.* **5,** 392–398 (1948).
9. Chance, B. Optical method. *Annu. Rev. Biophys. Biophys. Chem.* **20,** 1–28 (1991).
10. Jobsis, F. F. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science* **198,** 1264–1267 (1977).
11. Tatsuta, M. *et al.* Diagnosis of gastric cancers with fluorescein-labeled monoclonal antibodies to carcinoembryonic antigen. *Lasers Surg. Med.* **9,** 422–426 (1989).
12. Folli, S. *et al.* Immunophotodiagnosis of colon carcinomas in patients injected with fluoresceinated chimeric antibodies against carcinoembryonic antigen. *Proc. Natl Acad. Sci. USA* **89,** 7973–7977 (1992).
13. Pelegrin, A. *et al.* Antibody–fluorescein conjugates for photoimmunodiagnosis of human colon carcinoma in nude mice. *Cancer* **67,** 2529–2537 (1991).
14. Wagner, R. *Erläuterungstafeln zur Physiologie und Entwicklungsgeschichte* (Voss, Leipzig, 1839).
15. Halin, C., Rodrigo Mora, J., Sumen, C. & von Andrian, U. H. *In vivo* imaging of lymphocyte trafficking. *Annu. Rev. Cell Dev. Biol.* **21,** 581–603 (2005).
16. Flusberg, B. A. *et al.* Fiber-optic fluorescence imaging. *Nature Methods* **2,** 941–950 (2005).
17. Kiesslich, R., Goetz, M., Vieth, M., Galle, P. R. & Neurath, M. F. Confocal laser endoscopy for *in vivo* diagnosis of colorectal cancer. *Nature Clin. Pract. Oncol.* **4,** 480–490 (2007).
18. Yelin, D. *et al.* Three-dimensional miniature endoscopy. *Nature* **443,** 765 (2006).
19. Ntziachristos, V., Ripoll, J., Wang, L. V. & Weissleder, R. Looking and listening to light: the evolution of whole-body photonic imaging. *Nature Biotechnol.* **23,** 313–320 (2005).
20. Grimm, J. *et al.* Use of gene expression profiling to direct *in vivo* molecular imaging of lung cancer. *Proc. Natl Acad. Sci. USA* **102,** 14404–14409 (2005).
21. Zacharakis, G. *et al.* Volumetric tomography of fluorescent proteins through small animals *in vivo*. *Proc. Natl Acad. Sci. USA* **102,** 18252–18257 (2005).
22. Yun, S. H. *et al.* Comprehensive volumetric optical microscopy *in vivo*. *Nature Med.* **12,** 1429–1433 (2006).
   This paper describes the development of a fibre-optic imaging technique that is useful for diagnostic imaging of epithelial disease.
23. Zhang, H. F., Maslov, K., Stoica, G. & Wang, L. V. Functional photoacoustic microscopy for high-resolution and noninvasive *in vivo* imaging. *Nature Biotechnol.* **24,** 848–851 (2006).
24. Liu, Y. *et al.* Optical markers in duodenal mucosa predict the presence of pancreatic cancer. *Clin. Cancer Res.* **13,** 4392–4399 (2007).
25. Evans, C. L. *et al.* Chemical imaging of tissue *in vivo* with video-rate coherent anti-Stokes Raman scattering microscopy. *Proc. Natl Acad. Sci. USA* **102,** 16807–16812 (2005).
26. Gross, S. & Piwnica-Worms, D. Spying on cancer: molecular imaging *in vivo* with genetically encoded reporters. *Cancer Cell* **7,** 5–15 (2005).
27. Giepmans, B. N., Adams, S. R., Ellisman, M. H. & Tsien, R. Y. The fluorescent toolbox for assessing protein location and function. *Science* **312,** 217–224 (2006).
28. Ventura, A. *et al.* Restoration of p53 function leads to tumour regression *in vivo*. *Nature* **445,** 661–665 (2007).
29. Sharpless, N. E. & Depinho, R. A. The mighty mouse: genetically engineered mouse models in cancer drug development. *Nature Rev. Drug Discov.* **5,** 741–754 (2006).
30. Ray, P. *et al.* Noninvasive quantitative imaging of protein–protein interactions in living subjects. *Proc. Natl Acad. Sci. USA* **99,** 3105–3110 (2002).
31. Perroy, J., Pontier, S., Charest, P. G., Aubry, M. & Bouvier, M. Real-time monitoring of ubiquitination in living cells by BRET. *Nature Methods* **1,** 203–208 (2004).
32. Jares-Erijman, E. A. & Jovin, T. M. FRET imaging. *Nature Biotechnol.* **21,** 1387–1395 (2003).
33. Paulmurugan, R., Massoud, T. F., Huang, J. & Gambhir, S. S. Molecular imaging of drug-modulated protein–protein interactions in living subjects. *Cancer Res.* **64,** 2113–2119 (2004).
34. Kelly, K. *et al.* Targeted nanoparticles for imaging incipient pancreatic ductal adenocarcinoma. *PLoS Med.* (in the press).
35. Weissleder, R., Kelly, K., Sun, E. Y., Shtatland, T. & Josephson, L. Cell-specific targeting of nanoparticles by multivalent attachment of small molecules. *Nature Biotechnol.* **23,** 1418–1423 (2005).
   This study used libraries of nanoparticles to identify cell-specific targeting agents that can be used for *in vivo* labelling.
36. Kelly, K. A., Waterman, P. & Weissleder, R. *In vivo* imaging of molecularly targeted phage. *Neoplasia* **8,** 1011–1018 (2006).

37. Pittet, M. J. et al. In vivo imaging of T cell delivery to tumors after adoptive transfer therapy. Proc. Natl Acad. Sci. USA **104,** 12457–12461 (2007).

38. Bulte, J. W. Intracellular endosomal magnetic labeling of cells. Methods Mol. Med. **124,** 419–439 (2006).

39. Arbab, A. S., Liu, W. & Frank, J. A. Cellular magnetic resonance imaging: current status and future prospects. Expert Rev. Med. Devices **3,** 427–439 (2006).

40. Pittet, M. J., Swirski, F. K., Reynolds, F., Josephson, L. & Weissleder, R. Labeling of immune cells for in vivo imaging using magnetofluorescent nanoparticles. Nature Protocols doi:10.1038/nprot.2006.11 (2006).

41. Swirski, F. K. A near-infrared cell tracker reagent for multiscopic in vivo imaging and quantification of leukocyte immune responses. PLoS ONE **2,** e1075 (2007).

42. Boon, T., Coulie, P. G., Van den Eynde, B. J. & van der Bruggen, P. Human T cell responses against melanoma. Annu. Rev. Immunol. **24,** 175–208 (2006).

43. Romero, P. et al. Ex vivo staining of metastatic lymph nodes by class I major histocompatibility complex tetramers reveals high numbers of antigen-experienced tumor-specific cytolytic T lymphocytes. J. Exp. Med. **188,** 1641–1650 (1998).

44. Curiel, T. J. et al. Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival. Nature Med. **10,** 942–949 (2004).

45. Chen, M. L. et al. Regulatory T cells suppress tumor-specific CD8 T cell cytotoxicity through TGF-β signals in vivo. Proc. Natl Acad. Sci. USA **102,** 419–424 (2005).

46. Nagaraj, S. et al. Altered recognition of antigen is a mechanism of CD8[+] T cell tolerance in cancer. Nature Med. **13,** 828–835 (2007).

47. Yang, L. et al. Abrogation of TGFβ signaling in mammary carcinomas recruits Gr-1[+]CD11b[+] myeloid cells that promote metastasis. Cancer Cell **13,** 23–35 (2008).

48. Boissonnas, A., Fetler, L., Zeelenberg, I. S., Hugues, S. & Amigorena, S. In vivo imaging of cytotoxic T cell infiltration and elimination of a solid tumor. J. Exp. Med. **204,** 345–356 (2007).

49. Mrass, P. et al. Random migration precedes stable target cell interactions of tumor-infiltrating T cells. J. Exp. Med. **203,** 2749–2761 (2006).
   References 48 and 49 were the first in vivo studies of T-cell infiltration and motility in tumour stroma and T-cell formation of cognate-antigen-dependent contacts with tumour cells.

50. Veiga-Fernandes, H., Walter, U., Bourgeois, C., McLean, A. & Rocha, B. Response of naive and memory CD8[+] T cells to antigen stimulation in vivo. Nature Immunol. **1,** 47–53 (2000).

51. Mempel, T. R. et al. Regulatory T cells reversibly suppress cytotoxic T cell function independent of effector differentiation. Immunity **25,** 129–141 (2006).
   This study found that T$_{reg}$ cells reversibly suppress CTL-mediated antitumour immunity by allowing CTLs to acquire full effector potential but withholding their 'license to kill'.

52. Zippelius, A. et al. Effector function of human tumor-specific CD8 T cells in melanoma lesions: a state of local functional tolerance. Cancer Res. **64,** 2865–2873 (2004).

53. Wang, W., Eddy, R. & Condeelis, J. The cofilin pathway in breast cancer invasion and metastasis. Nature Rev. Cancer **7,** 429–440 (2007).

54. Wyckoff, J. B. et al. Direct visualization of macrophage-assisted tumor cell intravasation in mammary tumors. Cancer Res. **67,** 2649–2656 (2007).
   This paper reports that the intravasation of breast tumour cells in vivo occurs in association with perivascular TAMs.

55. Orimo, A. et al. Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. Cell **121,** 335–348 (2005).

56. Granot, D. et al. In vivo imaging of the systemic recruitment of fibroblasts to the angiogenic rim of ovarian carcinoma tumors. Cancer Res. **67,** 9180–9189 (2007).

57. Muller, A. et al. Involvement of chemokine receptors in breast cancer metastasis. Nature **410,** 50–56 (2001).

58. Soucek, L. et al. Mast cells are required for angiogenesis and macroscopic expansion of Myc-induced pancreatic islet tumors. Nature Med. **13,** 1211–1218 (2007).

59. Shojaei, F. et al. Bv8 regulates myeloid-cell-dependent tumour angiogenesis. Nature **450,** 825–831 (2007).

60. Karnoub, A. E. et al. Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. Nature **449,** 557–563 (2007).

61. Weissleder, R., Tung, C. H., Mahmood, U. & Bogdanov, A. Jr. In vivo imaging of tumors with protease-activated near-infrared fluorescent probes. Nature Biotechnol. **17,** 375–378 (1999).

62. Ntziachristos, V., Tung, C. H., Bremer, C. & Weissleder, R. Fluorescence molecular tomography resolves protease activity in vivo. Nature Med. **8,** 757–760 (2002).

63. Carmeliet, P. & Jain, R. K. Angiogenesis in cancer and other diseases. Nature **407,** 249–257 (2000).

64. McDonald, D. M. & Choyke, P. L. Imaging of angiogenesis: from microscope to clinic. Nature Med. **9,** 713–725 (2003).

65. Ferrara, N. & Kerbel, R. S. Angiogenesis as a therapeutic target. Nature **438,** 967–974 (2005).

66. Brown, E. B. et al. In vivo measurement of gene expression, angiogenesis and physiological function in tumors using multiphoton laser scanning microscopy. Nature Med. **7,** 864–868 (2001).
   This study used intravital microscopy and tumour-window-chamber models to monitor gene expression, cell adhesion, delivery of therapeutics, angiogenesis and blood-vessel permeability in deep regions of tumours.

67. Willett, C. G. et al. Direct evidence that the VEGF-specific antibody bevacizumab has antivascular effects in human rectal cancer. Nature Med. **10,** 145–147 (2004).

68. Montet, X., Ntziachristos, V., Grimm, J. & Weissleder, R. Tomographic fluorescence mapping of tumor targets. Cancer Res. **65,** 6330–6336 (2005).

69. Jain, R. K. Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. Science **307,** 58–62 (2005).

70. Contag, C. H., Jenkins, D., Contag, P. R. & Negrin, R. S. Use of reporter genes for optical measurements of neoplastic disease in vivo. Neoplasia **2,** 41–52 (2000).

71. Lee, K. C. et al. Noninvasive molecular imaging sheds light on the synergy between 5-fluorouracil and TRAIL/Apo2L for cancer therapy. Clin. Cancer Res. **13,** 1839–1846 (2007).

72. Laxman, B. et al. Noninvasive real-time imaging of apoptosis. Proc. Natl Acad. Sci. USA **99,** 16551–16555 (2002).

73. Ntziachristos, V. et al. Visualization of antitumor treatment by means of fluorescence molecular tomography with an annexin V–Cy5.5 conjugate. Proc. Natl Acad. Sci. USA **101,** 12294–12299 (2004).

74. Blankenberg, F. G., Vanderheyden, J. L., Strauss, H. W. & Tait, J. F. Radiolabeling of HYNIC–annexin V with technetium-99m for in vivo imaging of apoptosis. Nature Protocols doi:10.1038/nprot.2006.17 (2006).

75. Avril, N. et al. Prediction of response to neoadjuvant chemotherapy by sequential F-18-fluorodeoxyglucose positron emission tomography in patients with advanced-stage ovarian cancer. J. Clin. Oncol. **23,** 7445–7453 (2005).

76. Lordick, F. et al. PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: the MUNICON phase II trial. Lancet Oncol. **8,** 797–805 (2007).

77. Dekker, E. & Fockens, P. New imaging techniques at colonoscopy: tissue spectroscopy and narrow band imaging. Gastrointest. Endosc. Clin. N. Am. **15,** 703–714 (2005).

78. Herth, F. J., Eberhardt, R. & Ernst, A. The future of bronchoscopy in diagnosing, staging and treatment of lung cancer. Respiration **73,** 399–409 (2006).

79. Marten, K. et al. Detection of dysplastic intestinal adenomas using enzyme-sensing molecular beacons in mice. Gastroenterology **122,** 406–414 (2002).

80. Joyce, J. A. et al. Cathepsin cysteine proteases are effectors of invasive growth and angiogenesis during multistage tumorigenesis. Cancer Cell **5,** 443–453 (2004).
   This paper shows that cathepsins promote invasive growth and angiogenesis in pancreatic islet tumours.

81. McCarthy, J. R., Jaffer, F. A. & Weissleder, R. A macrophage-targeted theranostic nanoparticle for biomedical applications. Small **2,** 983–987 (2006).

82. Koyama, Y. et al. Spectral fluorescence molecular imaging of lung metastases targeting HER2/neu. Clin. Cancer Res. **13,** 2936–2945 (2007).

83. Kirsch, D. G. et al. A spatially and temporally restricted mouse model of soft tissue sarcoma. Nature Med. **13,** 992–997 (2007).
   This study combined a new mouse model of soft-tissue sarcoma that mimics human sarcomas with a hand-held imaging device that identifies residual tumour tissue during intra-operative molecular imaging.

84. Kim, S. et al. Near-infrared fluorescent type II quantum dots for sentinel lymph node mapping. Nature Biotechnol. **22,** 93–97 (2004).

85. Rudin, M. & Weissleder, R. Molecular imaging in drug discovery and development. Nature Rev. Drug Discov. **2,** 123–131 (2003).

86. Rontgen, W. C. On a new kind of rays. Nature **53,** 274–276 (1896).

87. Drexler, B., Davis, J. L. & Schofield, G. Diaphanography in the diagnosis of breast cancer. Radiology **157,** 41–44 (1985).

88. Figueiredo, J. L., Alencar, H., Weissleder, R. & Mahmood, U. Near infrared thoracoscopy of tumoral protease activity for improved detection of peripheral lung cancer. Int. J. Cancer **118,** 2672–2677 (2006).

89. Ntziachristos, V., Yodh, A. G., Schnall, M. & Chance, B. Concurrent MRI and diffuse optical tomography of breast after indocyanine green enhancement. Proc. Natl Acad. Sci. USA **97,** 2767–2772 (2000).

90. Drahl, C., Cravatt, B. F. & Sorensen, E. J. Protein-reactive natural products. Angew. Chem. Int. Ed. Engl. **44,** 5788–5809 (2005).

91. Roberti, M. J., Bertoncini, C. W., Klement, R., Jares-Erijman, E. A. & Jovin, T. M. Fluorescence imaging of amyloid formation in living cells by a functional, tetracysteine-tagged α-synuclein. Nature Methods **4,** 345–351 (2007).

92. Venkatraman, P. et al. Fluorogenic probes for monitoring peptide binding to class II MHC proteins in living cells. Nature Chem. Biol. **3,** 222–228 (2007).

93. Harisinghani, M. G. et al. Noninvasive detection of clinically occult lymph-node metastases in prostate cancer. N. Engl. J. Med. **348,** 2491–2499 (2003).

94. Cristofanilli, M. et al. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. N. Engl. J. Med. **351,** 781–791 (2004).

95. Georgakoudi, I. et al. In vivo flow cytometry: a new method for enumerating circulating cancer cells. Cancer Res. **64,** 5044–5047 (2004).
   This paper reports the first in vivo flow-cytometry-based method for quantifying circulating tumour cells.

96. Boutrus, S. et al. Portable two-color in vivo flow cytometer for real-time detection of fluorescently-labeled circulating cells. J. Biomed. Opt. **12,** 020507, doi:10.1117/1.2722733 (2007).

97. He, W., Wang, H., Hartmann, L. C., Cheng, J. X. & Low, P. S. In vivo quantitation of rare circulating tumor cells by multiphoton intravital flow cytometry. Proc. Natl Acad. Sci. USA **104,** 11760–11765 (2007).

98. Nagrath, S., Haber, D. A. & Toner, M. Isolation of rare circulating epithelial cells in cancer patients by microchip technology. Nature **450,** 1235–1239 (2007).

99. Lee, H., Sun, E. Y., Ham, D. & Weissleder, R. Chip–NMR biosensor for detection and molecular analysis of cells. Nature Med. (in the press).
   This study developed a low-cost diagnostic MRI platform for rapid, quantitative, multichannel detection of biological targets in unprocessed samples.

100. Perez, J. M., Josephson, L., O'Loughlin, T., Hogemann, D. & Weissleder, R. Magnetic relaxation switches capable of sensing molecular interactions. Nature Biotechnol. **20,** 816–820 (2002).

589

Sequence- and target-independent angiogenesis suppression by siRNA via TLR3

Mark E. Kleinman[1]*, Kiyoshi Yamada[1]*, Atsunobu Takeda[1]*, Vasu Chandrasekaran[3], Miho Nozaki[1], Judit Z. Baffi[1], Romulo J. C. Albuquerque[1,2], Satoshi Yamasaki[4], Masahiro Itaya[4], Yuzhen Pan[5], Binoy Appukuttan[5], Daniel Gibbs[6,7], Zhenglin Yang[6,7], Katalin Karikó[8], Balamurali K. Ambati[6,9], Traci A. Wilgus[10], Luisa A. DiPietro[10], Eiji Sakurai[4], Kang Zhang[6,7], Justine R. Smith[5], Ethan W. Taylor[11] & Jayakrishna Ambati[1,2]

Clinical trials of small interfering RNA (siRNA) targeting vascular endothelial growth factor-A (*VEGFA*) or its receptor *VEGFR1* (also called *FLT1*), in patients with blinding choroidal neovascularization (CNV) from age-related macular degeneration, are premised on gene silencing by means of intracellular RNA interference (RNAi). We show instead that CNV inhibition is a siRNA-class effect: 21-nucleotide or longer siRNAs targeting non-mammalian genes, non-expressed genes, non-genomic sequences, pro- and anti-angiogenic genes, and RNAi-incompetent siRNAs all suppressed CNV in mice comparably to siRNAs targeting *Vegfa* or *Vegfr1* without off-target RNAi or interferon-α/β activation. Non-targeted (against non-mammalian genes) and targeted (against *Vegfa* or *Vegfr1*) siRNA suppressed CNV via cell-surface toll-like receptor 3 (TLR3), its adaptor TRIF, and induction of interferon-γ and interleukin-12. Non-targeted siRNA suppressed dermal neovascularization in mice as effectively as *Vegfa* siRNA. siRNA-induced inhibition of neovascularization required a minimum length of 21 nucleotides, a bridging necessity in a modelled 2:1 TLR3–RNA complex. Choroidal endothelial cells from people expressing the TLR3 coding variant 412FF were refractory to extracellular siRNA-induced cytotoxicity, facilitating individualized pharmacogenetic therapy. Multiple human endothelial cell types expressed surface TLR3, indicating that generic siRNAs might treat angiogenic disorders that affect 8% of the world's population, and that siRNAs might induce unanticipated vascular or immune effects.

Therapeutic application of long, double-stranded (ds)RNA-mediated RNAi and sequence-specific gene silencing through RNAi by short synthetic RNA duplexes is challenging because mammalian cells do not uptake 'naked' siRNA (whether chemically modified or not) without cell-permeating entities[1–4]. To minimize systemic exposure, initial clinical trials of siRNA were launched using intraocular injection in patients with CNV. CNV, wherein the retina is invaded by choroidal vessels beneath the retinal pigmented epithelium (RPE), is a late stage of age-related macular degeneration that afflicts 30–50 million people globally. The preclinical bases for trials of naked *VEGFA* siRNA (Bevasiranib) or *VEGFR1* siRNA (AGN211745/ siRNA-027) were single reports in mice[5,6] that such siRNAs suppressed laser-injury-induced CNV, a model predictive of efficacy in humans[7,8]. These findings were interpreted as anomalous examples of local delivery surmounting the impediment to intracellular entry[9–11]. Instead, we show in two animal models that suppression of neovascularization is a generic property of siRNAs independent of sequence, target and internalization.

Sequence-independent angiogenesis suppression by siRNA

Numerous synthetic non-targeted 21-nucleotide duplex siRNAs from multiple vendors, when injected into the vitreous humour of wild-type mice, uniformly and dose-dependently suppressed CNV (Fig. 1a, b and Supplementary Fig. 1). siRNAs targeting jellyfish green fluorescent protein (*Gfp*) or firefly luciferase (*Luc*), non-ocular genes *Bglap1* (bone-specific osteocalcin), *Cdh16* (kidney-specific cadherin 16), or *Sftpb* (lung-specific surfactant protein B), or non-genomic random sequences (RS1–6), all suppressed CNV. This stereotypic effect, reproduced independently in the laboratories of J.A. and E.S., cannot be attributed to 'off-target' silencing due to sequence-specific mismatch tolerance[12], nor is it an artefact of intraocular delivery because intraperitoneal administration of serum-stable 2′O-methyl-*Luc*-siRNA suppressed CNV (Supplementary Fig. 2). Furthermore, a siRNA with random sequence and containing proprietary chemical modifications precluding incorporation into RNA-induced silencing complex (RISC) also suppressed CNV (Fig. 1a), whereas a dsDNA analogue of *Luc* siRNA did not (Supplementary Fig. 3). Lipopolysaccharide did not reduce CNV, excluding endotoxin contamination as the source of angio-inhibition; however, nuclease digestion did abolish *Luc*-siRNA-induced CNV suppression. Collectively, these data indicate that angio-inhibition is a siRNA-class effect.

Surface TLR3 mediates anti-angiogenic activity of siRNA

Consistent with reports that naked siRNAs are not internalized by mammalian cells[1–3], fluorescein-conjugated *Luc* siRNA did not enter

591

Figure 1 | Sequence-independent CNV suppression by siRNA through TLR3.
a, siRNAs targeting *Gfp*, *Luc*, random sequences (RS1, RS2), *Bglap1*, *Cdh16*, or
Sftpb, and siRNA incapable of RNA-induced silencing (RISC-free), suppressed
CNV in wild-type mice. *n* = 8–24; asterisk, *P* < 0.05 compared to no injection,
phosphate buffered saline (PBS) and siRNA buffer. **b**, Representative
examples of CNV in wild-type eye injected with vehicle (buffer) or *Luc* siRNA
(1 µg). **c**, *Gfp* siRNA or *Luc* siRNA suppressed CNV in *Tlr3*[+/+] but not *Tlr3*[−/−]
mice. *n* = 16–18; asterisk, *P* < 0.05 compared to vehicle (buffer).
d, Representative examples of CNV in *Gfp*-siRNA-injected (1 µg) *Tlr3*[+/+] and
Tlr3[−/−] eyes. **e**, CNV suppression in wild-type mice by *Luc* siRNA (0.25 µg)
was abrogated by soluble TLR3 (sTLR3) but not soluble TLR4 or heat-
denatured (hd) soluble TLR3 (all 2 µg). *n* = 8. **f**, CNV suppression in wild-type
mice by *Luc* siRNA (1 µg) was abrogated by neutralizing anti-TLR3 antibodies
(Ab; 0.2 µl) but not control IgG (0.2 µl) or chloroquine (Cq; 30 ng). *n* = 6–8;
asterisk, *P* < 0.05. NS, not significant. Vehicle, buffer. All error bars indicate
mean ± s.e.m. Scale bars in **b**, **d** are 100 µm.

primary human choroidal endothelial cells (CECs) or mouse RPE
and CECs (Supplementary Fig. 4). However, fluorescein–*Luc*-
siRNA–chol, in which cholesterol is conjugated to the 3′ end to
enable uptake[3], was internalized by human and mouse cells, both
of which express the mammalian homologue of Sid1, a protein
involved in internalizing cholesterol-conjugated siRNAs[13]. We
therefore sought a surface receptor mediating siRNA-induced
angio-inhibition.

We proposed that siRNAs activate TLR3, a long double-stranded
viral RNA sensor[14], to suppress CNV. Indeed, non-targeted siRNAs
did not suppress CNV in *Tlr3*[−/−] mice (Fig. 1c, d). The TLR3 acti-
vators polyinosinic:polycytidylic acid (poly(I:C))[14] and *in-vitro*-
transcribed long dsRNA[15] suppressed CNV in wild-type mice,
whereas the TLR3 non-activators polydeoxyinosinic:polydeoxy-
cytidylic acid (poly(dI:dC))[14] and 2-thiouridine (s2U)-modified *in-
vitro*-transcribed dsRNA[15] were ineffective. Poly(I:C) did not
suppress CNV in *Tlr3*[−/−] mice (Supplementary Fig. 5).

Soluble TLR3, but not soluble TLR4 or heat-denatured soluble
TLR3, abolished CNV suppression by *Luc* siRNA in wild-type mice
(Fig. 1e), suggesting direct interaction of siRNA with TLR3.
Fluorescein-conjugated *Luc* siRNA bound wild-type but not
Tlr3[−/−] mouse eye sections *in situ* (Supplementary Fig. 6). Using
flow cytometry to monitor binding of fluorescein–*Luc* siRNA to
the surface of CD31[+]VEGFR2[+] mouse choroidal endothelial cells,
greater fluorescence was detected on wild-type than *Tlr3*[−/−] cells
(Supplementary Fig. 7). Pre-incubation with soluble TLR3 or com-
petition with poly(I:C) reduced fluorescein–*Luc*-siRNA binding to
wild-type mouse CECs. These data extend cell-free system data that
20-nucleotide dsRNAs bind TLR3 (refs 16, 17) and support an *in vivo*
siRNA–TLR3 interaction, although we cannot exclude accessory
molecules enabling TLR3 activation. However, the only such faci-
litator reported so far, CD14 (ref. 18), was dispensable because *Luc*
siRNA suppressed CNV in *Cd14*[−/−] mice (Supplementary Fig. 8).

CNV suppression by *Luc* siRNA was blocked by TLR3-neutralizing
antibodies (Fig. 1f), suggesting that non-targeted siRNA signalled via
surface TLR3. We confirmed *in vivo* surface TLR3 expression on
mouse and human CECs by flow cytometry and immunofluore-
scence (Supplementary Fig. 9). To resolve the locus of TLR3 activa-
tion by siRNA, we used chloroquine, which inhibits endosomal TLR3
and TLR9. Chloroquine blocked the increase in CNV induced by
CpG oligonucleotide, a TLR9 agonist, but did not prevent CNV
suppression by *Luc* siRNA (Fig. 1f and Supplementary Fig. 10).
Collectively, these data show that surface, not endosomal, TLR3
mediates extracellular siRNA-induced angio-inhibition.

Non-targeted siRNAs did not suppress CNV in *Trif*[Lps2] mice
(Supplementary Fig. 11), which are deficient in signalling induced
by TRIF (toll/interleukin (IL)-1-receptor-domain-containing adaptor-
inducing interferon-β), the TLR3 adaptor protein[19,20], confirming
TLR3 indispensability. TLR3 signalling can diverge at the level of
TRIF through a kinase cascade activating interferon regulatory factor
(IRF)-3 or nuclear factor-κB (NF-κB)[21]. Non-targeted siRNAs
suppressed CNV in *Irf3*[−/−] but not *Nfkb1*[−/−] mice (Supplementary
Fig. 11), suggesting a TRIF–NF-κB cascade.

Duplex siRNAs containing immunostimulatory sequences such as
UGUGU[22] or GUCCUUCAA[23] trigger off-target effects via TLR7, a
single-stranded RNA sensor. However, none of our siRNAs con-
tained these sequences (Supplementary Table). Also, non-targeted
siRNAs suppressed CNV in *Tlr7*[−/−] mice (Supplementary Fig. 12).
Furthermore, 2′*O*-methyl-*Luc*-siRNA suppressed CNV, excluding
TLR7 involvement[24].

Although it is improbable that cytosolic dsRNA sensors such as
protein kinase R (PKR, encoded by *Prkra*), retinoic-acid-inducible
gene 1 (RIG-I), and melanoma differentiation-associated gene 5
(MDA5)[25–28] are involved in siRNA-mediated angio-inhibition
because *Luc* siRNA did not enter cells, we excluded their involvement
because siRNA activates PKR when transfected into cells[29]. *Luc*
siRNA suppressed CNV in *Prkra*[−/−] just as in wild-type mice,
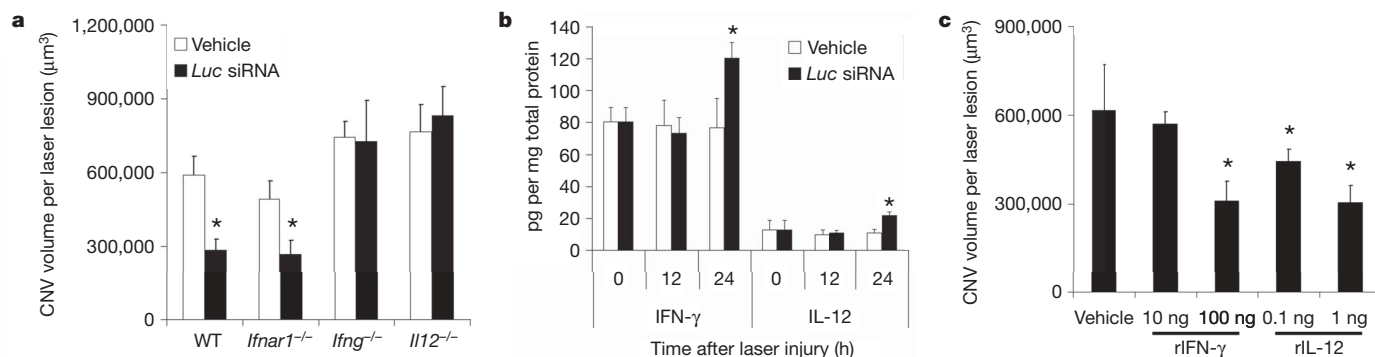
Figure 2 | CNV suppression by siRNA is mediated by IFN-γ and IL-12. a, *Luc* siRNA (1 μg) suppressed CNV in wild-type and *Ifnar1*$^{-/-}$ mice but not *Ifng*$^{-/-}$ or *Il12*$^{-/-}$ mice. $n = 5$–10. **b**, IFN-γ and IL-12 levels in RPE and choroid at 24 h after laser injury were higher in wild-type mouse eyes injected with *Luc* siRNA (1 μg) compared to vehicle-injected eyes. $n = 12$. **c**, Recombinant IFN-γ or IL-12 reduced CNV in wild-type mice. $n = 8$. Asterisk, $P < 0.05$ compared to vehicle (siRNA buffer). All error bars indicate mean ± s.e.m.

suggesting that PKR has no involvement (Supplementary Fig. 13). Poly(I:C$_{12}$U), an analogue of poly(I:C) that activates TLR3 but not MDA5 (ref. 30), suppressed CNV in wild-type mice as effectively as poly(I:C) (Supplementary Fig. 5). RIG-I participation is unlikely because *in-vitro*-transcribed dsRNA, which activates RIG-I, and poly(I:C), which does not[25], suppressed CNV equally (Supplementary Fig. 5). Thus, TLR3 seems to be central to siRNA-induced angio-inhibition.

Anti-angiogenic innate immunity triggered by siRNA

siRNAs sometimes induce type I interferon (IFN)[22,23,29,31–33]. Although laser injury increased levels of IFN-α and IFN-β (IFN-α/β) in the RPE and choroid, they were not further induced by *Luc* siRNA (Supplementary Fig. 14). *Luc* siRNA suppressed CNV in *Ifnar1*$^{-/-}$ (IFN-α/β receptor null) mice just as in wild-type mice (Fig. 2a). However, type II IFN-γ and IL-12, which are TLR3-inducible, were upregulated by *Luc* siRNA after laser injury (Fig. 2b). Recombinant IFN-γ and IL-12 suppressed CNV in wild-type mice (Fig. 2c), consistent with their anti-angiogenic properties[34]. Both were required for

angiogenesis suppression because *Luc* siRNA did not suppress CNV in *Ifng*$^{-/-}$ or *Il12a*$^{-/-}$ mice (Fig. 2a).

IFN-γ and IL-12 polarization is congruent with TRIF activation being biased towards NF-κB rather than IRF-3. Non-induction of IFN-α/β by synthetic siRNA is consistent with the absence of 5′-triphosphates[33] and the dominant role of MDA5, which is not involved in siRNA-induced angio-inhibition (Supplementary Fig. 5), in mediating dsRNA-induced IFN-α/β (refs 25, 28). Alternatively, laser injury might have saturated or inhibited IFN-α/β induction.

Targeted siRNAs also suppress CNV via TLR3

We synthesized a siRNA targeting the same sequence as AGN211745 in *Vegfr1* (*Vegfr1* siRNA, complementary to mouse and human sequences). *Vegfr1* siRNA suppressed CNV in wild-type mice, confirming an earlier report[6], but not in *Tlr3*$^{-/-}$ mice (Fig. 3a), suggesting that it suppressed angiogenesis by means of the class effect, not RNAi. Supportive of this, *Vegfr1* siRNA suppressed CNV as effectively as *Luc* siRNA in mice that lack the tyrosine kinase domain of *Vegfr1* (*Vegfr1*(*TK*)$^{-/-}$ mice, which have no VEGFR1 signalling[35];
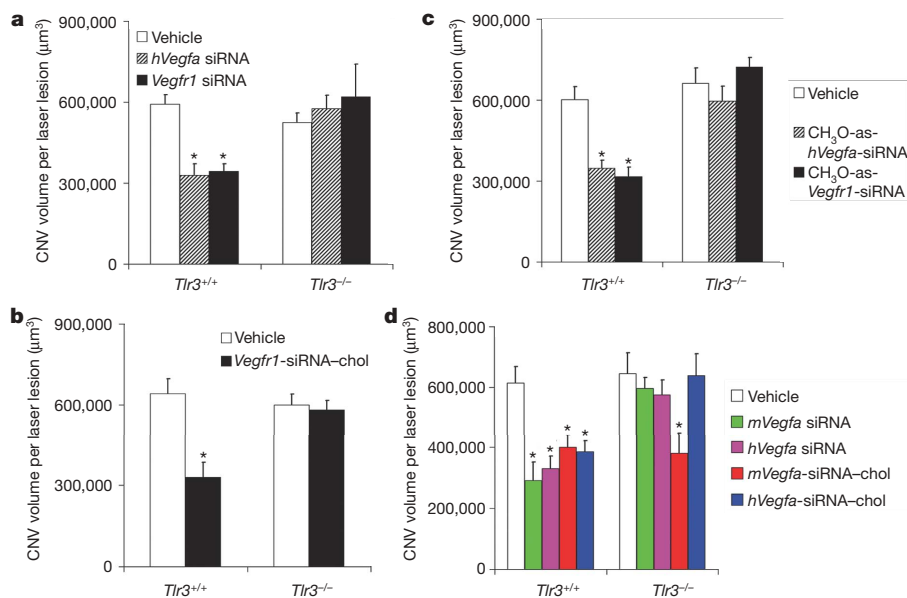
Figure 3 | CNV suppression by targeted siRNAs is mediated by TLR3. a, *hVegfa* siRNA and *Vegfr1* siRNA suppressed CNV in *Tlr3*$^{+/+}$ but not *Tlr3*$^{-/-}$ mice. **b**, *Vegfr1*-siRNA–chol suppressed CNV in *Tlr3*$^{+/+}$ but not *Tlr3*$^{-/-}$ mice. $n = 5$–6. **c**, *hVegfa* siRNA and *Vegfr1* siRNA, whose anti-sense (as) strands were modified with 5′-methoxy (CH$_3$O) substitution, preventing incorporation into RISC, suppressed CNV in *Tlr3*$^{+/+}$ but not *Tlr3*$^{-/-}$ mice. $n = 9$–14. **d**, siRNAs targeting mouse (m) or human (h) *Vegfa* reduced CNV in *Tlr3*$^{+/+}$ but not *Tlr3*$^{-/-}$ mice. *mVegfa*-siRNA–chol and *hVegfa*-siRNA–chol suppressed CNV in *Tlr3*$^{+/+}$ mice; however, *mVegfa*-siRNA–chol but not *hVegfa*-siRNA–chol suppressed CNV in *Tlr3*$^{-/-}$ mice. $n = 8$–10. siRNAs were at 1 μg; asterisk, $P < 0.05$ compared to vehicle (siRNA buffer). All error bars indicate mean ± s.e.m.

Supplementary Fig. 15). Also, *Vegfr1* siRNA did not suppress *Vegfr1* mRNA in the RPE and choroid of *Tlr3*$^{-/-}$ mice (Supplementary Fig. 15). Furthermore, *Vegfr1*-siRNA–chol did not suppress CNV in *Tlr3*$^{-/-}$ mice despite *Vegfr1* knockdown (Fig. 3b and Supplementary Fig. 15), consonant with observations that VEGFR1 does not promote CNV[36].

CH$_3$O-as-*Vegfr1* siRNA (containing 5′-methoxy modification of the anti-sense strand, preventing phosphorylation of the 5′-hydroxyl end obligate for RISC incorporation and target cleavage[37,38], thereby abolishing RNAi activity) suppressed CNV in wild-type mice as effectively as *Vegfr1* siRNA and was ineffective in *Tlr3*$^{-/-}$ mice (Fig. 3c), confirming that *Vegfr1* siRNA suppressed CNV through TLR3, not RNAi.

We also synthesized a siRNA targeting the same sequence as Bevasiranib in human *VEGFA* (*hVegfa* siRNA, complementary to human *VEGFA* but mismatched to mouse *Vegfa* at position 11, critical for functional targeting[39]). Both *hVegfa* siRNA and CH$_3$O-as-*hVegfa*-siRNA suppressed CNV in wild-type but not *Tlr3*$^{-/-}$ mice (Fig. 3a, c), demonstrating that *hVegfa* siRNA also suppresses CNV through TLR3, not RNAi. *hVegfa*-siRNA–chol did not suppress CNV in *Tlr3*$^{-/-}$ mice (Fig. 3d), indicating that this siRNA was incapable of functional targeting even after internalization due to critical nucleotide mismatch with mouse *Vegfa*. However, *mVegfa*-siRNA–chol (complementary to mouse *Vegfa*) suppressed CNV in *Tlr3*$^{-/-}$ mice. As further confirmation of the centrality of this nucleotide, *mVegfa*-siRNA–chol, but not *hVegfa*-siRNA–chol, suppressed *Vegfa* in the RPE and choroid of *Tlr3*$^{-/-}$ mice (Supplementary Fig. 16). Hence, siRNAs unaided by cell permeation cannot execute RNAi.

Luc siRNA suppressed CNV (51 ± 7%) in wild-type mice as effectively as *mVegfa*-siRNA–chol in *Tlr3*$^{-/-}$ mice (41 ± 11%; Fig. 3d) and anti-VEGFA neutralizing antibodies in wild-type mice (49 ± 7%; Supplementary Fig. 17), suggesting that TLR3 activation is as effective as VEGFA blockade, the current standard of care. *Luc* siRNA did not affect *Vegfa*, *Vegfr1*, or *Vegfr2* mRNA levels in wild-type mice (Supplementary Fig. 18), so activity seems to be independent of

VEGFA pathways. Despite the potential for *mVegfa*-siRNA–chol to activate TLR3 and knockdown *Vegfa* in wild-type mice, there was no angio-inhibitory synergy. This could be attributed to blockade of VEGFA-dependent signalling by IFN-γ (induced by siRNA) and potential crosstalk[40]. Indeed, CNV suppression by *Luc* siRNA or recombinant IFN-γ was not enhanced by co-administration of anti-VEGFA antibodies, consistent with convergence of anti-angiogenic mechanisms (Supplementary Fig. 18). Alternatively, incorporation of *mVegfa*-siRNA–chol into RISC might preclude TLR3 interaction.

Neither *hVegfa* siRNA nor *Vegfr1* siRNA suppressed CNV in *Ifng*$^{-/-}$ or *Il12a*$^{-/-}$ mice (Supplementary Fig. 19), confirming that these targeted siRNAs suppressed CNV via the class effect, not RNAi. As an additional demonstration that siRNA suppresses CNV without target cleavage, *Irf3* siRNA suppressed CNV in *Irf3*$^{-/-}$ mice as effectively as *Luc* siRNA (Supplementary Fig. 20), and *Tlr3* siRNA suppressed CNV in wild-type mice without *Tlr3* knockdown. This class effect was so robust that siRNA targeting *Il12*, which is anti-angiogenic in this model, also suppressed CNV in wild-type but not *Tlr3*$^{-/-}$ mice (Fig. 2c and Supplementary Fig. 21). These data demonstrate that non-targeted and targeted siRNA, without active intracellular delivery, suppressed CNV by means of extracellular TLR3 activation.

Minimum length of siRNA required to suppress CNV

The minimum length of dsRNA activating TLR3 *in vivo* is unclear. We found, unlike 21- and 23-nucleotide *Luc* siRNA, 7-, 13-, 16-, or 19-nucleotide versions did not suppress CNV (Fig. 4a). This, coupled with data that longer duplexes such as ~1,000-nucleotide dsRNA and poly(I:C) suppressed CNV (Supplementary Fig. 5), suggest that at least 21 nucleotides are required to activate TLR3. To understand the structural basis of the divergence between 19 and 21 nucleotides, we performed docking simulations between dimeric TLR3, which is required for signalling[17,41,42], and dsRNA using crystal structures of TLR3 and 19-nucleotide dsRNA, the latter being shortened or extended as necessary.
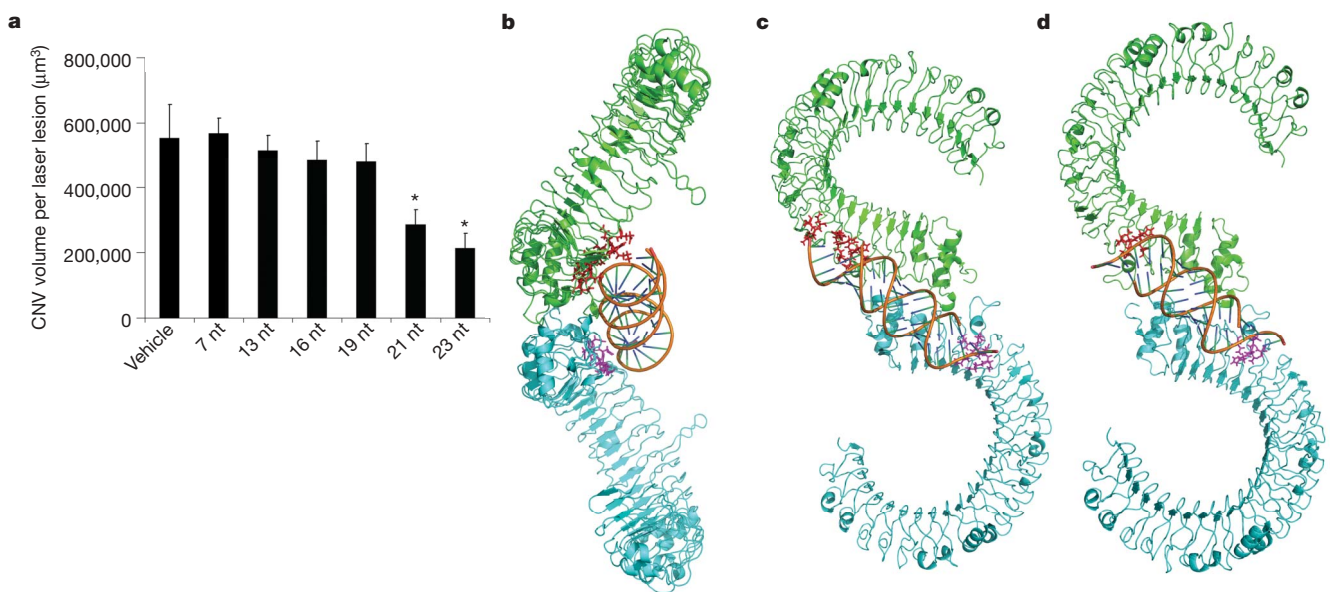
Figure 4 | Minimum length for TLR3 activation. a, 21-nucleotide or 23-nucleotide *Luc* siRNA but not truncated versions suppressed CNV in wild-type mice. $n = 8$–11; asterisk, $P < 0.05$ compared to vehicle (buffer); mean ± s.e.m. Equimolar amounts to 1 µg of 21-nucleotide *Luc* siRNA. nt, nucleotide. **b**, **c**, Orthogonal views of a model of TLR3 ectodomain dimer (green and cyan subunits shown as backbone ribbons) with computer-docked 21-nucleotide dsRNA. Protein modules interact via C-terminal domains to form a highly symmetrical dimer. The model incorporated

potential interactions involving a larger set of TLR3 residues considered important in RNA binding (displayed as purple and red side chains on ectodomain subunits); these residues were within 4.5 Å of RNA. **d**, TLR3 dimer with docked 19-nucleotide dsRNA shows proximity to fewer putative RNA-binding residues. Binding of 19-nucleotide dsRNA was less favourable than 21-nucleotide dsRNA ($\Delta E = -308$ kcal mol^{-1}; 95% confidence interval: 261–355; $n = 5$; $P = 0.008$).

594

Docking of dsRNA of various lengths on a TLR3 ectodomain (ECD) monomer did not clarify the observed minimal length requirement for TLR3 activation because the ECD solenoid is approximately the same width as a single turn of an RNA helix. Therefore, a TLR3 monomer cannot discriminate between dsRNAs greater than about 10 nucleotides in length (data not shown), indicating that TLR3 dimerization is required for the observed length needed for activation. We developed a model of a 2:1 TLR3–dsRNA complex, constrained by these observations or interpretations of known data: (1) active TLR3 involves a specific ECD carboxy terminus dimer, formation of which is stabilized by dsRNA binding[17,41,42]; (2) TLR3 residues most critical for RNA binding and receptor activation, H539 and N541 (ref. 41), lie in a region 40–45 Å from the ECD C terminus; (3) dsRNA of 21 nucleotides or longer is required for receptor activation (Fig. 4a), suggesting that a dsRNA of approximately two full helical turns (~50 Å in length) is required to span clusters of RNA-binding residues on two TLR3 monomers.

We proposed that C terminus dimerization of TLR3 ECDs creates a scaffold for RNA binding in which clusters of functional RNA-binding residues are separated by approximately 50 Å. This is possible with significant overlap of the C terminus domains, as would be required for effective dimerization.

Using these constraints, computer docking identified a highly symmetrical C terminus dimer of TLR3 ECD with clusters of RNA-binding residues on the same side of the complex and spaced appropriately (N541 to N541 distance of 50 Å). This dimer was used to dock either 21-nucleotide or 19-nucleotide dsRNA, under the constraint that docked complexes place H539 and N541 on both TLR3 monomers within 5 Å of RNA and be as symmetrical as possible[43]. The resulting 2:1 complexes (Fig. 4b–d) were subjected to short molecular dynamics simulations, from which five time points were sampled and energy minimized, indicating that the 21-nucleotide complex was approximately $300\,\text{kcal}\,\text{mol}^{-1}$ energetically more favourable than the 19-nucleotide complex. This is consistent with experimental results showing a loss in anti-angiogenic activity for dsRNA with fewer than 21-nucleotides (Fig. 4a). Furthermore, 19-nucleotide dsRNA was unable to make contact with as many of the putative RNA-binding residues as 21-nucleotide dsRNA (Fig. 4c, d and Supplementary Figs 22 and 23). The model also allows for binding of longer dsRNAs such as poly(I:C).

We considered possible interactions between dsRNA and an extended set of putative RNA-binding residues including, in addition to H539 and N541, E442, N466, K467, R489, N515, N516, N517, N540 and N572 (refs 17, 44). Compared to 19-nucleotide dsRNA, 21-nucleotide dsRNA made contact (within 4.5 Å) with 80% more side chains. Thus, this model accords with the possibility that these TLR3 residues may contribute in varying degrees to RNA binding, if not actual receptor activation.

This model predicts that although siRNAs shorter than 21 nucleotides cannot bridge functional residues on TLR3 dimers, they can bind just as well as longer siRNAs to TLR3 monomers (Fig. 4d and Supplementary Fig. 23). Indeed, excess 19-nucleotide *Luc* siRNA prevented CNV suppression by 21-nucleotide *Luc* siRNA (Supplementary Fig. 24) in a dose-dependent manner, suggesting that the shorter siRNA can be an inactive competitor by saturating binding sites on TLR3 monomers and interfere with the ability of longer siRNAs to bridge two monomers to form the active dimeric signalling complex. The specificity of this competition was demonstrated by the inability of a dsDNA analogue of 19-nucleotide *Luc* siRNA to interfere with CNV suppression by 21-nucleotide *Luc* siRNA.

Clinical implications

Of the known non-synonymous single-nucleotide polymorphisms (SNPs) in *TLR3*, two (N284I, rs5743316 and L412F, rs3775291) are hypomorphic (associated with a reduction in cell surface expression and functional activation of receptor)[45]. Variance for rs3775291 is prevalent in many populations (minor allele frequency ranges from 0.26 to 0.40 in Asian and European samples) whereas variance for rs5743316 is rare (minor allele frequency <0.01). Therefore we sought to determine whether non-targeted siRNA therapy would have differential effects on patients with different homozygous genotypes at rs3775291 (we did not identify any patients with variance at rs5743316). We scanned a library of primary human CECs isolated from single donors for cells homozygous at rs3775291. To determine the fate of these cells when exposed to extracellular non-targeted siRNA, we assayed their survival in serum-supplemented culture in the presence of serum-stable 2′O-methyl-*Luc*-siRNA without transfection reagents. Non-targeted siRNA induced a 38% reduction in survival of 412LL human CECs (consistent with known cytotoxic effects of TLR3 activation), but a 3% increase in that of 412FF human CECs (Supplementary Fig. 25), suggesting that the TLR3 412 genotype is an important determinant of the cell survival response to siRNA and consistent with TLR3-mediated angio-inhibition. However, further confirmation is required to assign differential survival solely to this SNP and to support a role for pharmacogenetic testing of patients before initiating siRNA therapeutics. Apart from CECs, pericytes and leukocytes have critical roles in angiogenesis; defining their contributions awaits the development of tissue-specific TLR3 knockouts.

To determine the potential for siRNA-induced angio-inhibition in other organs, we interrogated human endothelial cells from aorta, dermis, lung and umbilical vein for TLR3 expression. Surface TLR3 was detected on all of these cells (Supplementary Fig. 26). Serum-stable 21-nucleotide 2′O-methyl-*Luc*-siRNA, but not a 7-nucleotide version, suppressed dermal angiogenesis induced by excisional wounding in wild-type mice as effectively as 21-nucleotide 2′O-methyl-*mVegfa*-siRNA (Supplementary Fig. 27), recapitulating the length-dependent anti-angiogenic class effect of siRNA in CNV. The broad angio-inhibitory effect of siRNA and widespread surface expression of TLR3 on endothelial cells suggest that people with the 412LL genotype might be suitable candidates for exploring the anti-angiogenic utility of non-targeted siRNA, whereas targeted cell-permeating siRNA therapies would be better suited for 412FF individuals. Analysing SNPs of genes downstream from *TLR3* might yield refined genomic profiles predictive of response to targeted and non-targeted siRNAs.

Discussion

Our results demonstrate that 21-nucleotide or longer dsRNAs are anti-angiogenic. We show that two investigational siRNAs in clinical trials owe their anti-angiogenic effect in mice not to target knock-down but to TLR3 activation. Our findings support investigations of non-targeted dsRNAs as generic anti-CNV agents as effective as anti-VEGFA antibodies, the current standard of care[8], while avoiding potential neurotoxicity resulting from chronic administration of the latter[46].

Our docking model, although not a high-resolution structure, confirms a potential interaction between 21-nucleotide siRNA and TLR3, and structurally explains TLR3's remarkable length-based discrimination of dsRNAs. Although duplexes shorter than 21 nucleotides did not suppress CNV, our model suggests that they could interact with TLR3 with a free energy of binding below the threshold for dimer stabilization and receptor activation. Such discrimination between interaction and activation, if it exists, might also be attributed to unknown accessory proteins that chaperone siRNA to TLR3. An outstanding question is whether differential binding to monomeric versus active signalling dimeric TLR3 underlies the divergence between 21-nucleotide and shorter siRNAs.

Tlr3$^{+/+}$ and *Tlr3*$^{-/-}$ mice exhibited similar CNV responses in the absence of exogenous dsRNAs, suggesting that TLR3 was dispensable to basal angiogenesis and that endogenous dsRNAs (for example, microRNAs) do not activate TLR3 or influence angiogenesis in this

model. Elucidating the discriminants among potential TLR3 ligands will advance understanding of self/non-self immune recognition.

It is tempting to speculate that TLR3 activation and anti-angiogenic dampening by viral transcripts abundant in the human retina–choroid interface (G. S. Hageman, personal communication) might explain why few patients with early age-related macular degeneration undergo the angiogenic switch despite the onslaught of inflammatory stimuli that promote CNV[47]. Similarly, investigations of SNPs in *TLR3* and downstream genes might prove fruitful in revealing new risk-conferring and protective alleles for CNV and atrophic age-related macular degeneration, and predict efficacy of angiogenesis inhibition and susceptibility to adverse vascular effects.

Further studies will determine whether siRNAs, apart from inducing (un)desirable effects on the vasculature, also influence TLR3-mediated antimicrobial immune responses and non-canonical adverse effects of TLR3 activation on pregnancy[48], immune privilege[49] and neuronal growth[50], especially when delivered systemically. Judicious approaches to enhance therapeutic specificity include modifications abrogating TLR3 activation and reducing lengths below 21 nucleotides. Conversely, genomic fingerprinting of patients could hasten angio-inhibitory applications of generic siRNA while avoiding hazards of specific gene targeting.

METHODS SUMMARY

Laser photocoagulation was performed to induce CNV in mice as previously described[36,47]. CNV volumes were measured by scanning laser confocal microscopy using fluorescein-isothiocyanate-conjugated *Griffonia simplicifolia* isolectin B4 (Vector Laboratories), and compared by hierarchical logistic regression using repeated measures analysis. Chemical siRNA sequences for treatments were purchased from Dharmacon or Ambion, Inc. Sense and anti-sense strands of siRNA were annealed per the manufacturer's instructions, formulated in either sterile siRNA buffer (Dharmacon) or nuclease-free PBS, and injected into the vitreous cavity in a total volume of 1 μl with a 33-gauge Exmire microsyringe (Ito Corporation). The docked complex structures were obtained through rigid-body docking using a modified version of the program FTDock with the X-ray structures of TLR3 ligand-binding domain (Protein Data Bank (PDB) entry 2A0Z) and a 19-nucleotide RNA duplex (PDB entry 1QC0) as the initial structures, the latter being shortened or extended as necessary to generate various lengths of dsRNA with 2-nucleotide overhangs at each end. Final docked 2:1 TLR3–dsRNA complexes were subjected to a short minimization and an implicit solvent molecular dynamics simulation to refine the structures, which were then used for energy calculations. Application of biological information as described above greatly improved the docking results by constraining the sampling space and generating models consistent with the experimental binding data.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Chiu, Y. L., Ali, A., Chu, C. Y., Cao, H. & Rana, T. M. Visualizing a correlation between siRNA localization, cellular uptake, and RNAi in living cells. *Chem. Biol.* **11**, 1165–1175 (2004).
2. Saleh, M. C. *et al.* The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing. *Nature Cell Biol.* **8**, 793–802 (2006).
3. Soutschek, J. *et al.* Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature* **432**, 173–178 (2004).
4. Song, E. *et al.* Antibody mediated *in vivo* delivery of small interfering RNAs via cell-surface receptors. *Nature Biotechnol.* **23**, 709–717 (2005).
5. Reich, S. J. *et al.* Small interfering RNA (siRNA) targeting VEGF effectively inhibits ocular neovascularization in a mouse model. *Mol. Vis.* **9**, 210–216 (2003).
6. Shen, J. *et al.* Suppression of ocular neovascularization with siRNA targeting VEGF receptor 1. *Gene Ther.* **13**, 225–234 (2006).
7. Krzystolik, M. G. *et al.* Prevention of experimental choroidal neovascularization with intravitreal anti-vascular endothelial growth factor antibody fragment. *Arch. Ophthalmol.* **120**, 338–346 (2002).
8. Rosenfeld, P. J. *et al.* Ranibizumab for neovascular age-related macular degeneration. *N. Engl. J. Med.* **355**, 1419–1431 (2006).
9. Corey, D. R. RNA learns from antisense. *Nature Chem. Biol.* **3**, 8–11 (2007).
10. de Fougerolles, A., Vornlocher, H. P., Maraganore, J. & Lieberman, J. Interfering with disease: a progress report on siRNA-based therapeutics. *Nature Rev. Drug Discov.* **6**, 443–453 (2007).
11. Rana, T. M. Illuminating the silence: understanding the structure and function of small RNAs. *Nature Rev. Mol. Cell Biol.* **8**, 23–36 (2007).
12. Jackson, A. L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnol.* **21**, 635–637 (2003).
13. Wolfrum, C. *et al.* Mechanisms and optimization of *in vivo* delivery of lipophilic siRNAs. *Nature Biotechnol.* **25**, 1149–1157 (2007).
14. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF-κB by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).
15. Kariko, K., Buckstein, M., Ni, H. & Weissman, D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* **23**, 165–175 (2005).
16. Bell, J. K. *et al.* The molecular structure of the Toll-like receptor 3 ligand-binding domain. *Proc. Natl Acad. Sci. USA* **102**, 10976–10980 (2005).
17. Ranjith-Kumar, C. T. *et al.* Biochemical and functional analyses of the human Toll-like receptor 3 ectodomain. *J. Biol. Chem.* **282**, 7668–7678 (2007).
18. Lee, H. K., Dunzendorfer, S., Soldau, K. & Tobias, P. S. Double-stranded RNA-mediated TLR3 activation is enhanced by CD14. *Immunity* **24**, 153–163 (2006).
19. Yamamoto, M. *et al.* Role of adaptor TRIF in the MyD88-independent toll-like receptor signaling pathway. *Science* **301**, 640–643 (2003).
20. Hoebe, K. *et al.* Identification of Lps2 as a key transducer of MyD88-independent TIR signalling. *Nature* **424**, 743–748 (2003).
21. Jiang, Z., Mak, T. W., Sen, G. & Li, X. Toll-like receptor 3-mediated activation of NF-κB and IRF3 diverges at Toll-IL-1 receptor domain-containing adapter inducing IFN-β. *Proc. Natl Acad. Sci. USA* **101**, 3533–3538 (2004).
22. Judge, A. D. *et al.* Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nature Biotechnol.* **23**, 457–462 (2005).
23. Hornung, V. *et al.* Sequence-specific potent induction of IFN-α by short interfering RNA in plasmacytoid dendritic cells through TLR7. *Nature Med.* **11**, 263–270 (2005).
24. Judge, A. D., Bola, G., Lee, A. C. & MacLachlan, I. Design of noninflammatory synthetic siRNA mediating potent gene silencing *in vivo*. *Mol. Ther.* **13**, 494–505 (2006).
25. Kato, H. *et al.* Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* **441**, 101–105 (2006).
26. Meurs, E. *et al.* Molecular cloning and characterization of the human double-stranded RNA-activated protein kinase induced by interferon. *Cell* **62**, 379–390 (1990).
27. Yoneyama, M. *et al.* The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nature Immunol.* **5**, 730–737 (2004).
28. Gitlin, L. *et al.* Essential role of mda-5 in type I IFN responses to polyriboinosinic:polyribocytidylic acid and encephalomyocarditis picornavirus. *Proc. Natl Acad. Sci. USA* **103**, 8459–8464 (2006).
29. Sledz, C. A., Holko, M., de Veer, M. J., Silverman, R. H. & Williams, B. R. Activation of the interferon system by short-interfering RNAs. *Nature Cell Biol.* **5**, 834–839 (2003).
30. Gowen, B. B. *et al.* TLR3 is essential for the induction of protective immunity against Punta Toro Virus infection by the double-stranded RNA (dsRNA), poly(I:C12U), but not Poly(I:C): differential recognition of synthetic dsRNA molecules. *J. Immunol.* **178**, 5200–5208 (2007).
31. Heidel, J. D., Hu, S., Liu, X. F., Triche, T. J. & Davis, M. E. Lack of interferon response in animals to naked siRNAs. *Nature Biotechnol.* **22**, 1579–1582 (2004).
32. Kariko, K., Bhuyan, P., Capodici, J. & Weissman, D. Small interfering RNAs mediate sequence-independent gene suppression and induce immune activation by signaling through toll-like receptor 3. *J. Immunol.* **172**, 6545–6549 (2004).
33. Kim, D. H. *et al.* Interferon induction by siRNAs and ssRNAs synthesized by phage polymerase. *Nature Biotechnol.* **22**, 321–325 (2004).
34. Voest, E. E. *et al.* Inhibition of angiogenesis *in vivo* by interleukin 12. *J. Natl Cancer Inst.* **87**, 581–586 (1995).
35. Hiratsuka, S., Minowa, O., Kuno, J., Noda, T. & Shibuya, M. Flt-1 lacking the tyrosine kinase domain is sufficient for normal development and angiogenesis in mice. *Proc. Natl Acad. Sci. USA* **95**, 9349–9354 (1998).
36. Nozaki, M. *et al.* Loss of SPARC-mediated VEGFR-1 suppression after injury reveals a novel antiangiogenic activity of VEGF-A. *J. Clin. Invest.* **116**, 422–429 (2006).
37. Nykanen, A., Haley, B. & Zamore, P. D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
38. Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R. & Tuschl, T. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**, 563–574 (2002).
39. Birmingham, A. *et al.* 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods* **3**, 199–204 (2006).
40. Battle, T. E., Lynch, R. A. & Frank, D. A. Signal transducer and activator of transcription 1 activation in endothelial cells is a negative regulator of angiogenesis. *Cancer Res.* **66**, 3649–3657 (2006).
41. Bell, J. K., Askins, J., Hall, P. R., Davies, D. R. & Segal, D. M. The dsRNA binding site of human Toll-like receptor 3. *Proc. Natl Acad. Sci. USA* **103**, 8792–8797 (2006).
42. Choe, J., Kelker, M. S. & Wilson, I. A. Crystal structure of human toll-like receptor 3 (TLR3) ectodomain. *Science* **309**, 581–585 (2005).
43. Gay, N. J., Gangloff, M. & Weber, A. N. Toll-like receptors as molecular switches. *Nature Rev. Immunol.* **6**, 693–698 (2006).
44. Takada, E. *et al.* C-terminal LRRs of human Toll-like receptor 3 control receptor dimerization and signal transmission. *Mol. Immunol.* **44**, 3633–3640 (2007).

45. Ranjith-Kumar, C. T. *et al.* Effects of single nucleotide polymorphisms on Toll-like Receptor 3 activity and expression in cultured cells. *J. Biol. Chem.* **282**, 17696–17705 (2007).

46. Nishijima, K. *et al.* Vascular endothelial growth factor-A is a survival factor for retinal neurons and a critical neuroprotectant during the adaptive response to ischemic injury. *Am. J. Pathol.* **171**, 53–67 (2007).

47. Nozaki, M. *et al.* Drusen complement components C3a and C5a promote choroidal neovascularization. *Proc. Natl Acad. Sci. USA* **103**, 2328–2333 (2006).

48. Lin, Y., Liang, Z., Chen, Y. & Zeng, Y. TLR3-involved modulation of pregnancy tolerance in double-stranded RNA-stimulated NOD/SCID mice. *J. Immunol.* **176**, 4147–4154 (2006).

49. Lang, K. S. *et al.* Immunoprivileged status of the liver is controlled by Toll-like receptor 3 signaling. *J. Clin. Invest.* **116**, 2456–2463 (2006).

50. Cameron, J. S. *et al.* Toll-like receptor 3 is a potent negative regulator of axonal growth in mammals. *J. Neurosci.* **27**, 13033–13041 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions M.E.K., K.Y., A.T., M.N., J.Z.B., R.J.C.A., S.I., M.I. and E.S. performed CNV experiments. T.A.W. and L.A.D. performed dermal experiments. V.C. and E.W.T. performed modelling simulations. Y.P., B.A. and J.R.S. performed cytotoxicity experiments. D.G., Z.Y. and K.Z. performed genotyping. K.K. provided reagents. J.A. conceived and directed the project, and, with assistance from E.W.T., B.K.A., K.K., B.A. and J.R.S., wrote the paper. All authors had the opportunity to discuss the results and comment on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.A. (jamba2@email.uky.edu).

METHODS

Animals. C57BL/6J, *Ifng*$^{-/-}$, *Il12a*$^{-/-}$, *Nfkb1*$^{-/-}$, *Tlr3*$^{-/-}$, *Tlr3*$^{+/+}$ and *Trif*^{Lps2} mice were purchased from The Jackson Laboratory. *Vegfr1(TK)*$^{-/-}$, *Ifnar1*$^{-/-}$ (gift from H. W. Virgin), *Irf3*$^{-/-}$ (gift from T. Taniguchi via M. David), *Prkra*$^{-/-}$ (gift from R. H. Silverman via G. Luo) and *Tlr7*$^{-/-}$ (obtained from S. Akira via D. A. Golenbock) mice have been previously described[35,51-53].

CNV. Laser photocoagulation (OcuLight GL, IRIDEX) was performed on both eyes of each 6–8-week-old mice to induce CNV as previously described[36,47]. CNV volumes were measured by scanning laser confocal microscopy (TCS SP, Leica) as reported[36,47] with 0.5% FITC-conjugated Isolectin B4 (Vector). Results are expressed as mean ± s.e.m. with $P < 0.05$ considered statistically significant.

Drug treatments. siRNAs formulated in siRNA buffer (Dharmacon) or PBS (Sigma-Aldrich), chloroquine (Invivogen), poly(I:C) (Invivogen), poly(dI:dC) (Invivogen), poly(I:C$_{12}$U) (Ampligen; Bioclones), and recombinant soluble mouse TLR3 (R&D Systems), mouse soluble TLR4 (R&D Systems), neutralizing rat antibodies against mouse TLR3 (eBioscience; ref. 48), immunostimulatory CpG oligodeoxynucleotide (ODN), GpC ODN (Invivogen), recombinant IFN-γ (Peprotech), recombinant IL-12 (eBioscience), neutralizing rat antibodies against mouse IL-12 (R&D Systems) and isotype control IgGs (R&D Systems or eBioscience as appropriate) were dissolved in PBS and injected into the vitreous cavity in a total volume of 1 μl. In some experiments, siRNA or vehicle was injected intraperitoneally.

siRNA. Chemical siRNA sequences and modifications are listed in the Supplementary Table. All siRNAs except RS3–RS6 (Ambion) were purchased from Dharmacon.

Protein expression. ELISAs were used to quantify IFN-α/β/γ (PBL) with normalization to total protein (Bio-Rad).

Gene expression. Real-time reverse transcriptase PCR (RT–PCR) was performed (Realplex, Eppendorf) on total mouse RPE and choroid RNA (RNAqueous, Ambion). Primers and Taqman probes for *Vegfa* were described in ref. 54. SYBR green was used for all other assays. Primers for TLR3 were purchased (Superarray), and were custom designed for *Vegfr1* and *Vegfr2* (Supplementary Methods).

Flow cytometry. For surface TLR3 staining, human CECs, isolated as previously reported[55], human aortic endothelial cells (HAECs; gift from E. Reed), human pulmonary artery endothelial cells (HPAECs; Clonetics), human dermal microvascular endothelial cells (HDMECs; Clonetics) and human umbilical vein endothelial cells (HUVECs; Clonetics) cultivated in MCDB-131 media (Gibco) with EGM-2MV (omitting gentamicin, hydrocortisone, Clonetics) were blocked and incubated with PE-conjugated anti-human TLR3 (20 μg ml^{-1}, Imgenex) and FITC-conjugated CD31 (BD Biosciences). For intracellular TLR3 staining, TLR–CD31-labelled cells were fixed and permeabilized with Leucoperm (Serotec) and stained for TLR3 in the presence of 10% mouse serum. PE-conjugated mouse IgGκ1 isotype served as control (BD Biosciences). Samples were analysed on a FACSCalibur (Becton Dickinson) with a minimum of 10,000 events using Kolmogorov–Smirnov statistics.

Suspensions of C57BL/6J RPE and choroid cells (10^6) were isolated 1 day after laser injury and incubated with APC anti-mouse CD31 antibody (20 μg ml^{-1}, BD Biosciences) and PE anti-mouse VEGFR2 antibody (20 μg ml^{-1}; BD Biosciences). For surface TLR3 staining, cells were incubated with anti-mouse TLR3 (10 μg ml^{-1}, R&D Systems) that was pre-conjugated to Alexa Fluor 488 (Invitrogen). For intracellular TLR3 staining, CD31/VEGFR2-labelled cells were subjected to fixation and permeabilization followed by incubation with anti-mouse TLR3 (5 μg ml^{-1}). FITC-conjugated rat IgG2a isotype (BD Biosciences) served as a control.

Fluorescein–siRNA interaction assays. CD31/VEGFR2-labelled RPE and choroid cells (10^6) isolated from C57BL/6J or *Tlr3*$^{-/-}$ mice 1 day after laser injury were incubated with fluorescein–*Luc*-siRNA (100 μg ml^{-1}) for 30 min on ice. To determine the specificity *Luc* siRNA binding to TLR3, fluorescein–*Luc*-siRNA was pre-incubated with either soluble TLR3, TLR4, poly(I:C), or poly(dI:dC).

Fluorescein–siRNA uptake in mouse RPE and choroid cells. C57BL/6J RPE and choroid cells (10^6) were harvested immediately after laser injury, incubated with 21-nucleotide fluorescein–*Luc*-siRNA, 21-nucleotide fluorescein–*Luc*-siRNA–chol (both 100 μg ml^{-1}) or PBS vehicle in phenol-free M199 media (Gibco) plus 10% FBS for 1 h at 37 °C and treated with 150 μl of 0.05% trypsin (Gibco) for 3 min at 37 °C to remove extracellular siRNA.

Immunofluorescent microscopy. Human CECs (10^5) were incubated with 21-nucleotide fluorescein–*Luc*-siRNA or 21-nucleotide fluorescein–*Luc*-siRNA–chol (both 100 μg ml^{-1}) for 1 h at 37 °C.

Fluorescein–siRNA binding in mouse RPE and choroid tissue. Frozen sections of C57BL/6J or *Tlr3*$^{-/-}$ mice prepared 3 days after laser injury and intravitreous injection of *Luc* siRNA (1 μg) were blocked with 3% BSA and incubated with 21-nucleotide fluorescein–*Luc*-siRNA (10 μg ml^{-1}) for 1 h at 37 °C. Slides were fixed with 4% PFA, blocked in 10% normal rabbit serum, incubated with rabbit anti-FITC antibody (1:100, Zymed) overnight at 4 °C followed by anti-rabbit Alexa Fluor 488 (1:200, Invitrogen).

TLR3 localization. C57BL/6J mouse eye sections (10 μm) were fixed in ice-cold acetone for 10 min at 4 °C, blocked, and incubated with rabbit anti-mouse TLR3 (1:50, Imgenex) overnight at 4 °C followed by Alexa Fluor 594 goat anti-rabbit IgG (1:200, Invitrogen) For co-labelling of CECs, sections were incubated with rat anti-mouse CD31 (1:50, BD Biosciences) followed by goat anti-rat IgG conjugated to Alexa Fluor 488 (1:200). Human CECs, isolated from donated eyes using anti-CD31 antibody-coated magnetic beads (BD Biosciences and Invitrogen, respectively) as previously described[55], were grown in chamber slides (Lab-Tek). For surface staining, cells were blocked with PBS-1% BSA and incubated with mouse anti-human TLR3 (20 μg ml^{-1}, eBioscience). For surface membrane co-localization, cells were stained with rabbit anti-CD46 (1:50, Santa Cruz). Goat anti-mouse Alexa Fluor 488 and donkey anti-rabbit Cy3 (both 1:400) were used as secondary antibodies. Mouse IgGκ1 isotype antibody (20 μg ml^{-1}, eBioscience) served as a control.

51. Muller, U. *et al.* Functional role of type I and type II interferons in antiviral defense. *Science* **264**, 1918–1921 (1994).
52. Sato, M. *et al.* Distinct and essential roles of transcription factors IRF-3 and IRF-7 in response to viruses for IFN-α/β gene induction. *Immunity* **13**, 539–548 (2000).
53. Hemmi, H. *et al.* Small anti-viral compounds activate immune cells via the TLR7 MyD88-dependent signaling pathway. *Nature Immunol.* **3**, 196–200 (2002).
54. Zhang, L. *et al.* Different effects of glucose starvation on expression and stability of VEGF mRNA isoforms in murine ovarian cancer cells. *Biochem. Biophys. Res. Commun.* **292**, 860–868 (2002).
55. Smith, J. R. *et al.* Unique gene expression profiles of donor-matched human retinal and choroidal vascular endothelial cells. *Invest. Ophthalmol. Vis. Sci.* **48**, 2676–2684 (2007).

ARTICLES

Following translation by single ribosomes one codon at a time

Jin-Der Wen[1], Laura Lancaster[2], Courtney Hodges[3], Ana-Carolina Zeri[4], Shige H. Yoshimura[5], Harry F. Noller[2], Carlos Bustamante[1,3,6] & Ignacio Tinoco Jr[1]

We have followed individual ribosomes as they translate single messenger RNA hairpins tethered by the ends to optical tweezers. Here we reveal that translation occurs through successive translocation-and-pause cycles. The distribution of pause lengths, with a median of 2.8 s, indicates that at least two rate-determining processes control each pause. Each translocation step measures three bases—one codon—and occurs in less than 0.1 s. Analysis of the times required for translocation reveals, surprisingly, that there are three substeps in each step. Pause lengths, and thus the overall rate of translation, depend on the secondary structure of the mRNA; the applied force destabilizes secondary structure and decreases pause durations, but does not affect translocation times. Translocation and RNA unwinding are strictly coupled ribosomal functions.

Current understanding of the ribosome and the mechanism of translation has been significantly strengthened and expanded by recent advances in crystallography[1–6] and cryo-electron microscopy[7–10]. The ribosome undergoes several dynamical structural changes as it moves relative to the mRNA and transfer RNAs during translation[8,11]. Kinetic experiments have given a quantitative description of some of these dynamics during the main steps of the elongation cycle of protein synthesis[12]. During elongation, the secondary structures present in all mRNAs are disrupted to allow movement of the mRNA through the 30S subunit, and the reading of each codon. This task is aided by the mRNA helicase activity of the ribosome that has been localized to the downstream tunnel of the 30S subunit[13]. Moreover, interactions of mRNA pseudoknots or hairpins with the helicase region of the ribosome can shift the reading frame of the mRNA to the −1 frame, and play an important role in regulating gene expression in retroviruses[14–16].

It is extremely difficult to follow the steps of ribosomes during translational elongation using ensemble methods, because the dynamics of individual ribosomes are stochastic[17,18] and it is impossible to synchronize their activity. Here, we have used optical tweezers to follow the step-by-step translation of a single hairpin-forming mRNA molecule by a single ribosome. This approach has allowed us to characterize the dynamics of ribosome translation, measuring the time the ribosome spends at each codon, the number of mRNA nucleotides that move through the ribosome in each translocation step, and the time required per step. We have also determined the effects of mRNA structure on step size and rate, and have studied the effects of internal Shine–Dalgarno sequences[19] on translation arrest. These experiments provide a dynamic picture of the movement of a messenger RNA through a ribosome.

In these experiments, we used a single mRNA hairpin with a ribosome stalled at the 5′ end by omission of a required aminoacyl-tRNA; the RNA was attached to two micrometre-sized beads by RNA–DNA handles. One of the beads was held on a micropipette and the other in an optical trap used to measure the changes in distance between the beads (in nanometres) and the forces applied to the hairpin (in

piconewtons) (Fig. 1A)[20–22]. Translation is resumed at the single-molecule level by adding a mixture containing the required aminoacyl-tRNAs. During translation, the ribosome opens the hairpin as it moves through the RNA; thus, each base translocated requires the breaking of a base pair, which corresponds to an increase in the end-to-end distance of the mRNA by about 1 nm at the forces involved in these experiments (15–20 pN)[23]. Translation can thus be followed in real time by monitoring these changes in distance.

To establish that translation occurred in a single molecule of RNA held in the optical tweezers apparatus, we used a 60-base-pair (bp) hairpin that contains unique codons where the ribosome can be stalled (S3hp; see a in Fig. 1B). The progress of translation was established by determining the size of the residual hairpin, through its mechanical unfolding (details presented in Supplementary Information and Supplementary Fig. 1). To verify that the change in hairpin size was caused by a stalled ribosome at the corresponding position on the hairpin, we flowed a ribosome-releasing mixture containing puromycin (see Methods) into the reaction chamber. After a few minutes, the hairpin fully refolded, indicating that the ribosome had been released. These control experiments established that an RNA duplex can be unwound by the ribosome as translation proceeds, and that a ribosome can be stably stalled on the RNA at a chosen position.

We also translated the mRNA in bulk in the absence of force (using *Escherichia coli* S100 enzymes, see Supplementary information). The 60-bp S3hp mRNA and its corresponding wild-type ribosomal S3 mRNA were translated completely, but the hairpin RNA was translated approximately 40% slower (see Supplementary Fig. 2). These results show that ribosomes can translate long helices as used in our experiments, even without the aid of force to unfold them.

Following translation in real time

To follow translation by single ribosomes in real time, we designed another 60-bp hairpin construct, VE60hp, whose first 15 codons in the hairpin region encode only valine (Val) and glutamic acid (Glu), preceded by two phenylalanine codons (Phe) to allow ribosome stalling (see b in Fig. 1B). The hairpin with a ribosome stalled at the Phe

[1]Department of Chemistry, University of California, Berkeley, California 94720, USA. [2]Department of Molecular, Cell, and Developmental Biology, and Center for Molecular Biology of RNA, University of California, Santa Cruz, California 95064, USA. [3]Biophysics Graduate Group, University of California, Berkeley, California 94720, USA. [4]Brazilian Synchrotron Light Laboratory, Caixa Postal 6192, Campinas SP 13083-970, Brazil. [5]Graduate School of Biostudies, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan. [6]Howard Hughes Medical Institute, Department of Physics and Molecular and Cell Biology, University of California, Berkeley, California 94720, USA.

Figure 1 | Experimental design. A, The ribosome was stalled at the 5′ side of the mRNA hairpin construct, which was then held between two polystyrene beads coated with anti-digoxigenin antibody (top; approximately 3 μm in diameter) and streptavidin (bottom; approximately 2 μm). The micropipette was substituted by a second laser trap when VE274hp was used for translation. Drawings are schematic and not to scale. **B**, All the mRNA constructs used in this report share the same sequence and design except for the hairpin (upper case) and flanking regions (lower case). The Shine–Dalgarno sequence (AGGAG) and AUG start codon are highlighted in cyan, the tetraloop of the hairpin in green, Val codon patches in grey and Glu in yellow. The codons chosen to stall the ribosome are boxed. **a**, S3hp, 60 bp. **b**, VE60hp, 60 bp. **c**, VE274hp, 274 bp. The complementary sequence for VE274hp is not shown.

codon was held at a constant force of 20 pN—a force below that required to unfold the hairpin (approximately 23 pN, see Supplementary Fig. 3), and the extension change was monitored in time. As expected, hairpin unwinding was not observed before a translation mixture was injected into the reaction chamber to resume elongation. The extension and force trajectories versus time obtained in the presence of the translation mixture are shown in Fig. 2a. Strikingly, the extension shows a repeated step–pause–step pattern. These patterns display clear and discrete transitions (indicated by arrows). The steps were quantitatively analysed by calculating the corresponding pairwise distance distributions[23], as shown in Fig. 2b. The distance between steps is 2.7 nm, which corresponds to six nucleotides at

Figure 2 | Codon-by-codon translation of VE60hp. a, Extension and force trajectories during translation. The data were collected at 200 Hz (blue traces) and smoothed to 10 Hz (red). Discrete steps are indicated by arrowheads. The 18 nm rip at 163 s corresponds to spontaneous opening of the remaining approximately 18-bp hairpin ahead of the translating ribosome. **b**, Pairwise distance analysis of the extension trajectory in **a** from 147 to 157 s after correction for drift. **c**, The pattern of the translation process illustrated by a close-up view of the 148–150 s region. **d**, Distribution of translocation times. The translocation times ($n = 121$) were binned in 0.025 s intervals, and the probability density plotted and fitted to three possible Poisson equations (see Supplementary Information): one step (black line), two steps (blue) and three steps (red), with R^2 values of 0.45, 0.87 and 0.95, respectively. The k value from the best fit (red line) is $40 \pm 4\,\text{s}^{-1}$. Attempts to fit the distribution to two or three steps with different rate constants gave equal rate constants as the best fit.

20 pN; that is, the breaking of three hairpin base pairs accompanying a translocation step of three bases. Thus, we identified these transitions as ribosomal steps corresponding to the translation of individual codons. Figure 2c shows a step–pause–step trajectory defining the translocation time, the dwell time (the time between successive steps) and the pause time (the time between translocation events) of one such cycle. Direct analysis of the translocation events (number of events, $n = 121$) during translation of VE60hp gives a step size of 2.94 ± 0.72 bases occurring in 0.078 ± 0.045 s.

Although in our assay we directly observe the opening of hairpin base pairs and the release of the single-strand chains due to the helicase activity of the ribosome, and not its actual translocation, it is noteworthy that the ribosome opens exactly three base pairs, corresponding to the translation of each codon. The most parsimonious interpretation of this observation is that unwinding and translocation are strictly coupled functions and occur simultaneously; that is, within 0.078 s. Accordingly, we refer here to the transitions observed as translocations. These results constitute the first direct real-time observation, to our knowledge, of the physical steps of the ribosome machine moving along the mRNA during translation.

Translation of a 274-bp hairpin

We used a third mRNA construct (VE274hp; see c in Fig. 1C) harbouring a 274-bp hairpin with patches of Val and Glu codons to improve the statistics of translation dynamics. To follow translation on this longer messenger, we used a dual-trap optical tweezer instrument in which the pipette holding the bottom bead shown in Fig. 1A was replaced by a second optical trap. This instrument provided lower drift that improved data collection during the longer times required to translate this mRNA. However, as expected, the longer RNA also introduced more high-frequency noise.

Translation of VE274hp began by stalling the ribosome at the first Glu codon (E20; see c in Fig. 1B) of the hairpin. The RNA was held at an initial force of 17–18 pN before flowing the translation mixture (containing purified Val-tRNA^{Val} and Glu-tRNA^{Glu}) into the chamber to resume elongation. As the reaction proceeded, opening of the hairpin by the translating ribosome resulted in an increase in RNA extension, and a corresponding decrease in force[24], because the two laser beams of the dual-trap tweezers were held fixed during the experiments (Supplementary Fig. 4).

As seen in Fig. 3a, translation dynamics is a complex process, which is different for each reaction. The four translation trajectories shown in Fig. 3a were obtained for different ribosomes under the same reaction conditions, but the patterns vary significantly. Apart from runs of active translation, these trajectories also show variable overall translation rates, the existence of long pauses of variable lengths and locations, and positions where translation stopped altogether. Nonetheless, close-up views of active translation trajectories from different reactions (Fig. 3b) show the characteristic translocation–pause–translocation pattern with three-base steps, or multiples of them.

Only a small fraction of ribosomes ($n = 4$ out of 29) reach the last Glu patch (labelled E5; see, for example, Fig. 3a, blue trace) of the VE274hp. Most trajectories end earlier; we define these early halting events as 'translation arrests'. Figure 3c summarizes the positions of translation arrests in the VE274hp construct (underlying blue bars). One-third (9 out of 29) of the ribosomes stopped at a region (positions 67–75) that occurs just downstream from a Shine–Dalgarno-like AGGAGG sequence (positions 56–61, Fig. 3c). Moreover, another internal Shine–Dalgarno sequence (positions 167–172) also seems to favour translational arrests. To test the effects of internal Shine–Dalgarno sequences on translation, the first Shine–Dalgarno
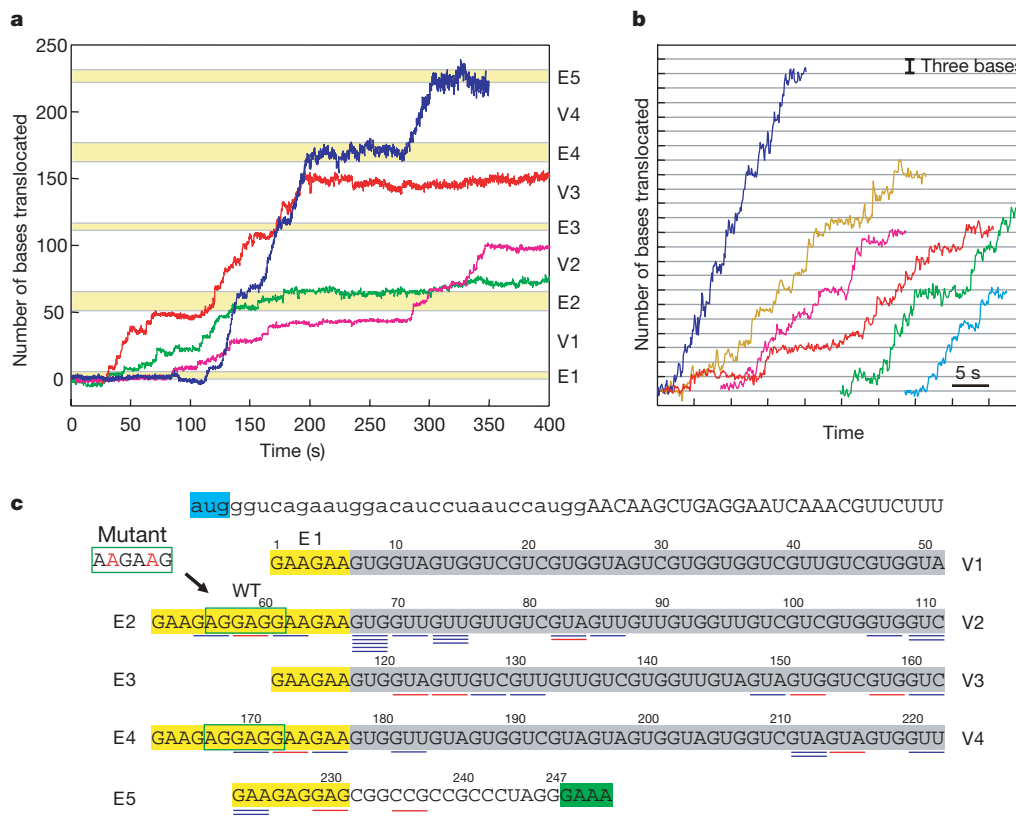
Figure 3 | Translational trajectories and arrests for VE274hp. Translations were followed on dual-trap optical tweezers[35]. **a**, Examples of translation trajectories from four different ribosomes. The five horizontal yellow stripes correspond to the Glu patches (see **c**). **b**, Close-up views of trajectories. Data were collected at 2 kHz and smoothed to 10 Hz. **c**, Translation arrests. The positions (relative to the A-site on the ribosome) where translation arrests occur are indicated by short horizontal bars, blue for the wild-type VE274hp and red for the mutant hairpin VE274hp-G2A. The VE274hp-G2A differs from the wild type only in the first internal Shine–Dalgarno-like sequence (boxed by solid green line).

sequence A<u>G</u> GA<u>G</u> G in VE274hp was changed to A<u>A</u> GA<u>A</u> G; note that the new codons are recognized by the same isotype of tRNA[Glu] and the amino-acid sequence is maintained. On this mutant hairpin, VE274hp-G2A, the translation arrest downstream of the Shine–Dalgarno sequence (positions 67–75) has now disappeared (underlying red bars, Fig. 3c). These results suggest that translating ribosomes can be stalled by base pairing between the 3′ end of 16S rRNA and internal Shine–Dalgarno sequences in the mRNA, similar to the interaction during the initiation process. A recent report shows that the Shine–Dalgarno interaction increases the rupture force of the ribosome from the bound mRNA by 10 pN[25]. The arrests observed in the vicinity of the Shine–Dalgarno sequence in our experiments indicate that this extra binding energy, together with the downstream hairpin structure, is sufficient to block the movement of the ribosome.
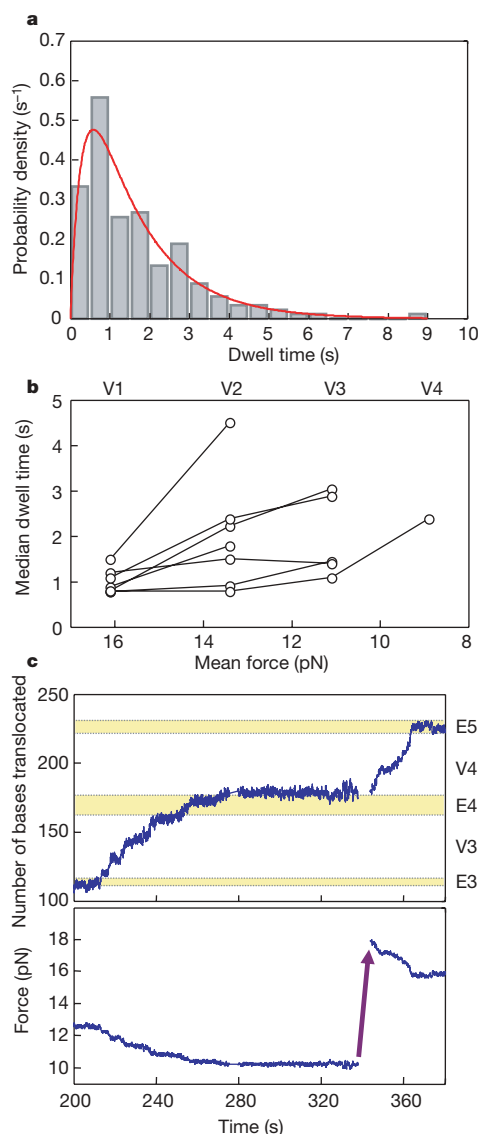
Figure 4 | Dwell times and force effect. a, Distribution of dwell times from VE274hp. Dwell-time data ($n = 180$) are pooled from translations of the first Val patch (V1) for 12 ribosomes with mean dwell times less than 2 s. The distribution (0.5 s bins) is well fitted to a mechanism in which two consecutive reactions take place (red line), with $k_1 = 0.7 \pm 0.2\,\text{s}^{-1}$, $k_2 = 3.4 \pm 1.8\,\text{s}^{-1}$ and $R^2 = 0.93$. **b**, Force effect on dwell times. The median dwell times in each Val patch are plotted as a function of force (mean force in each Val patch, V1–V4). Data points from the same ribosomes are connected by lines. **c**, Force effect on long pauses. In this example, the translation paused when the force dropped to about 10 pN. After approximately 1 min of pausing, the force was rapidly raised to 18 pN, and the translation resumed immediately.

Pauses and translocations

As shown in Fig. 3a, active translation runs made up of translocation–pause–translocation cycles are eventually interrupted by longer pauses, which can last up to 1–2 min (Supplementary Fig. 5). The long pauses occur less often in the first valine patch (V1) than in succeeding ones (Supplementary Fig. 6a). These long pauses can be identified from the derivatives of the individual translation trajectories (that is, the derivatives of the traces in Fig. 3a; one is shown in Supplementary Fig. 6b). The valleys between peaks on such plots characterize the long pauses, whereas the peaks on the plot correspond to active translation runs. These peaks show a relatively constant rate of 0.45 ± 0.17 codons per second, corresponding to a mean dwell time of 2.2 s ($n = 43$ ribosomes); these short pauses, we surmise, represent the constitutive (intrinsic) pauses of a translating ribosome.

The kinetics of translation have been studied in bulk[12], and in single-molecule fluorescence resonance energy transfer (FRET) experiments[17,18]; the rates of the rate-determining (slowest) steps are of order 1–$10\,\text{s}^{-1}$. In single-molecule experiments, the distribution of lifetimes for a species depends on the number and rates of the molecular events that transform that species into another. If a single stochastic event determines the lifetime, the distribution of lifetimes follows a single exponential decay from the smallest lifetimes to the longest lifetimes. For two or more successive reactions the distribution has a maximum at intermediate values of the lifetimes (see Supplementary Information)[26,27].

We analysed the dwell times—the times between successive steps—during translation of the first 15 valines in the VE274hp sequence. We found that these times for an actively translating subset of ribosomes (mean dwell time approximately 2 s) are not exponentially distributed. Instead, the data ($n = 180$) can be fitted well ($R^2 = 0.93$) to a difference of two exponentials (Fig. 4a). To ensure that our choice of ribosomes had not biased the results, we also analysed different subsets, and all the ribosomes. A difference of two exponentials was needed to fit each distribution, although the values of the rate constants, k, differed for each distribution. These results indicate that two, or more, successive rate-limiting processes control the dwell times between steps.

By contrast, and surprisingly, the distribution of translocation times for VE60hp best fits an equation (see Supplementary Information) with three identical substeps, with a mean lifetime of each of the three substeps of 0.025 ± 0.002 s ($k = 40\,\text{s}^{-1}$) (Fig. 2d). This result indicates that multiple processes with similar kinetics underlie the translocation step of three nucleotides. Until now, no technique has allowed characterization of the dynamics of ribosome stepping to this degree of detail. The kinetic processes we identified here could involve, for example, steps in the unlocking rearrangement of the ribosome[28]; or it is also possible that the ribosome moves over the length of one codon with three consecutive, equally timed, single-base substeps.

The effect of force on translation rates

Figure 3a shows that translation dynamics (overall rates, see Supplementary Fig. 7, and extents of reactions) are different for each ribosome. However, recurrence of Val patches on VE274hp allowed us to investigate, on the same ribosome, the effect of the force applied to the hairpin. Figure 4b shows that the lengths of pauses increase—translation rates decrease—when the force dropped as translation proceeded from Val segment 1 to 4 (V1–V4). The distributions of dwell times show that decreasing the force (16 to 9 pN) decreases the faster rate constant of the two substeps by a factor of three. At lower forces, the energy barrier impeding strand separation increases, resulting in slower translation rates and longer dwell times. This suggests that a paused ribosome could be rescued by raising the force applied on the ends of the hairpin. This is indeed the case, as shown in Fig. 4c, where raising the force from 10 pN to 18 pN caused translation to immediately resume. Similar force-rescued translations

were observed in about one-third of the long-paused ribosomes (14 events in 40 trials). Thus, some of the longer pauses observed in these experiments may be accounted for by the mechanical barriers imposed on the ribosome by the secondary structure of the mRNA.

Surprisingly, the force applied to the hairpin affects only the pause times; there is no detectable change in the translocation times. This observation implies that the rate-limiting step for translocation in our experiments corresponds to an intermediate state of the mRNA–tRNA–ribosome complex whose attainment is prevented, in part, by an activation barrier that is lowered by the force applied to the ends of the RNA. Once the complex has cleared this barrier, helix opening (and translocation) occurs rapidly and independently of force.

Biological implications

We have successfully applied optical tweezers to follow translation by single ribosomes in real time along individual molecules of mRNA. These studies have revealed for the first time that translation occurs not in a continuous manner, but as a series of translocation–pause–translocation events. These intrinsic short pauses, displayed by actively translating ribosomes, have durations that range from a fraction of a second to a few seconds; their mean is 2.2 s. In addition, we observe pauses that are significantly longer, lasting from tens of seconds up to 1–2 min. By contrast, the translocation steps, lasting for less than 0.1 s, were independent of the experimental variables we studied. The ability to measure the time spent at each codon (dwells) and the time required to move to the next codon (translocations) will make it possible to correlate biochemical results obtained in bulk studies with the dynamics of translation at a specific step or substep.

The ubiquitous long pauses observed in these studies may play a role in translation regulation *in vivo*, and can lead to translational frameshifting[29] and to protein misfolding[30]. A recent cryo-electron microscopy study on frameshifting suggests that mRNA pseudoknots cause conformational changes in the ribosome that create tension on the mRNA, which in turn results in −1 frameshifting[16]. Pauses caused by the increased barriers as ribosomes move through a hairpin[31], as in these experiments, can also lead to frameshifting.

We find that internal Shine–Dalgarno sequences tend to stop translation, presumably by binding to 16S rRNA during long pauses. The ability to measure dwell times at each codon opens the way to understanding the mechanism of translational regulation involving internal Shine–Dalgarno sequences, such as in the expression of the *prfB* and *dnaX* genes in *E. coli*[29]. The mRNA transcripts of these genes both contain internal Shine–Dalgarno sequences, and can lead to +1 or −1 frameshifts.

The studies on hairpins described here mimic biologically relevant situations of translation. Messenger RNAs are expected to have roughly half of their bases paired[32,33], mostly in short hairpins (5–10 bp), but more stable structures, including pseudoknots, are not uncommon. Future efforts should concentrate on single-molecule translation using less structured templates similar to naturally occurring mRNAs, and on measuring the translation signal directly by tethering the ribosome as it moves along the mRNA. By forming tethers[34] to the ribosome and the end of the linear mRNA, it will be possible to investigate the motor properties of the ribosome: that is, its stall force and its force–velocity relation. These properties can provide further insight into the mechanism of energy transduction of this essential cellular machine.

METHODS SUMMARY

Preparation of stalling translation mixtures. The stalling translation mixtures were used to stall the ribosome at designated positions on the mRNA. In general, the following components were mixed in 50 µl of Buffer TL: 2 mM ATP, 1 mM GTP, 5 mM PEP (phosphoenol pyruvate), 1 OD total tRNA mixture, 24 µM EF-Tu, 4 µM EF-G, 0.02 mg ml^{-1} pyruvate kinase (Roche), 1 µl DEAE-purified S-100 enzymes, 40 U RNAguard (RNase inhibitor; Amersham) and amino-acid mixture containing 0.1 mM of each amino acid. Buffer TL contained: 40 mM HEPES-KOH, pH 7.5, 60 mM NH$_4$Cl, 10 mM Mg(OAc)$_2$, 1 mM DTT and 3.6 mM 2-mercaptoethanol. The components of the amino-acid mixture

depended on the experiment. For example, to stall the ribosome at the arginine R11 of S3hp, the amino acids used were G, Q, K, V, H, P, N, and I (see a in Fig. 1B). The 50-µl mixture was incubated at 37 °C for 15 min and on ice for 5 min, and then diluted with three volumes of Buffer TL containing 1 mM GTP. **Preparation of initiation and stalled ribosome–mRNA complexes.** To make the initiation complex, the mRNA hairpin constructs (0.2 µM) were mixed with 1 mM GTP, 4 µM each of IF1, IF2, IF3, fMet-tRNA^{fMet}, and the 70S ribosome in 10 µl Buffer TL. The mixture was incubated at 37 °C for 15 min. Elongation was performed by mixing 1 µl initiation complex and 9 µl stalling translation mixture and incubating at 37 °C for 2–5 min. The reaction was used to stall the ribosome at the designated position on the mRNA. Finally, the mixture was diluted with 20 volumes of cold Buffer TL to prepare it for tweezer experiments.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Moore, P. B. & Steitz, T. A. The structural basis of large ribosomal subunit function. *Annu. Rev. Biochem.* **72**, 813–850 (2003).
2. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74**, 129–177 (2005).
3. Selmer, M. *et al.* Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **313**, 1935–1942 (2006).
4. Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).
5. Schuwirth, B. S. *et al.* Structures of the bacterial ribosome at 3.5 Å resolution. *Science* **310**, 827–834 (2005).
6. Korostelev, A., Trakhanov, S., Laurberg, M. & Noller, H. F. Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell* **126**, 1065–1077 (2006).
7. Frank, J. Electron microscopy of functional ribosome complexes. *Biopolymers* **68**, 223–233 (2003).
8. Frank, J. & Agrawal, R. K. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **406**, 318–322 (2000).
9. Allen, G. S., Zavialov, A., Gursky, R., Ehrenberg, M. & Frank, J. The cryo-EM structure of a translation initiation complex from *Escherichia coli*. *Cell* **121**, 703–712 (2005).
10. Stark, H. *et al.* Ribosome interactions of aminoacyl-tRNA and elongation factor Tu in the codon-recognition complex. *Nature Struct. Biol.* **9**, 849–854 (2002).
11. Horan, L. H. & Noller, H. F. Intersubunit movement is required for ribosomal translocation. *Proc. Natl Acad. Sci. USA* **104**, 4881–4885 (2007).
12. Wintermeyer, W. *et al.* Mechanisms of elongation on the ribosome: dynamics of a macromolecular machine. *Biochem. Soc. Trans.* **32**, 733–737 (2004).
13. Takyar, S., Hickerson, R. P. & Noller, H. F. mRNA helicase activity of the ribosome. *Cell* **120**, 49–58 (2005).
14. Alam, S. L., Atkins, J. F. & Gesteland, R. F. Programmed ribosomal frameshifting: much ado about knotting! *Proc. Natl Acad. Sci. USA* **96**, 14177–14179 (1999).
15. Brierley, I., Digard, P. & Inglis, S. C. Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell* **57**, 537–547 (1989).
16. Namy, O., Moran, S. J., Stuart, D. I., Gilbert, R. J. & Brierley, I. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature* **441**, 244–247 (2006).
17. Blanchard, S. C., Kim, H. D., Gonzalez, R. L. Jr, Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101**, 12893–12898 (2004).
18. Blanchard, S. C., Gonzalez, R. L., Kim, H. D., Chu, S. & Puglisi, J. D. tRNA selection and kinetic proofreading in translation. *Nat. Struct. Mol. Biol.* **11**, 1008–1014 (2004).
19. Shine, J. & Dalgarno, L. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA* **71**, 1342–1346 (1974).
20. Tinoco, I. Jr, Li, P. T. X. & Bustamante, C. Determination of thermodynamics and kinetics of RNA reactions by force. *Q. Rev. Biophys.* **39**, 325–360 (2006).
21. Liphardt, J., Onoa, B., Smith, S. B., Tinoco, I. Jr & Bustamante, C. Reversible unfolding of single RNA molecules by mechanical force. *Science* **292**, 733–737 (2001).
22. Onoa, B. *et al.* Identifying kinetic barriers to mechanical unfolding of the *T. thermophila* ribozyme. *Science* **299**, 1892–1895 (2003).
23. Dumont, S. *et al.* RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature* **439**, 105–108 (2006).
24. Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of lambda-phage DNA. *Science* **265**, 1599–1600 (1994).
25. Uemura, S. *et al.* Peptide bond formation destabilizes Shine–Dalgarno interaction on the ribosome. *Nature* **446**, 454–457 (2007).
26. Bustamante, C., Chemla, Y. R., Forde, N. R. & Izhaky, D. Mechanical processes in biochemistry. *Annu. Rev. Biochem.* **73**, 705–748 (2004).
27. Tinoco, I. Jr. Force as a useful variable in reactions: unfolding RNA. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 363–385 (2004).

28. Savelsbergh, A. *et al.* An elongation factor G-induced ribosome rearrangement precedes tRNA-mRNA translocation. *Mol. Cell* 11, 1517–1523 (2003).

29. Farabaugh, P. J. Programmed translational frameshifting. *Microbiol. Rev.* 60, 103–134 (1996).

30. Kimchi-Sarfaty, C. *et al.* A ''silent'' polymorphism in the *MDR1* gene changes substrate specificity. *Science* 315, 525–528 (2007).

31. Ivanov, I. P. & Atkins, J. F. Ribosomal frameshifting in decoding antizyme mRNAs from yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying conservation. *Nucleic Acids Res.* 35, 1842–1858 (2007).

32. Doty, P., Boedtker, H., Fresco, J. R., Haselkorn, R. & Litt, M. Secondary structure in ribonucleic acids. *Proc. Natl Acad. Sci. USA* 45, 482–499 (1959).

33. Favre, A., Morel, C. & Scherrer, K. The secondary structure and poly(A) content of globin messenger RNA as a pure RNA and in polyribosome-derived ribonucleoprotein complexes. *Eur. J. Biochem.* 57, 147–157 (1975).

34. Vanzi, F., Takagi, Y., Shuman, H., Cooperman, B. S. & Goldman, Y. E. Mechanical studies of single ribosome/mRNA complexes. *Biophys. J.* 89, 1909–1919 (2005).

35. Moffitt, J. R., Chemla, Y. R., Izhaky, D. & Bustamante, C. Differential detection of dual traps improves the spatial resolution of optical tweezers. *Proc. Natl Acad. Sci. USA* 103, 9006–9011 (2006).

METHODS

General. Configuration of the single-trap tweezers has been described previously[36,37]. Configuration of the dual-trap tweezers was similar to that described by Moffitt *et al.*[35], except that a 1064-nm laser was used. The *E. coli* MRE600 ribosomes[38], initiation factors (IF1, IF2 and IF3)[39], EF-G[40], EF-Tu[41] and S-100 enzymes[42] were purified as per protocols described in the literature. The ribosomes and proteins were made in small aliquots and stored at −80 °C.

Construction of plasmids. The plasmids used in the experiments were originally constructed from pBluescript SK+ (Stratagene). A 769-bp *E. coli* ribosomal S3 gene[43] was inserted between *Xba*I and *Not*I sites, and a 619-bp PCR fragment (from pBR322) was inserted into the *Kpn*I site, between the T7 promoter sequence and the S3 gene fragment, to serve as the 5′ handle. To make pS3hp (for S3hp mRNA, see a in Fig. 1B), a 107-bp dsDNA was inserted between *Nco*I and *Bsr*GI sites (inside the S3 gene), such that the sequence for codons 13–32 of S3 was base-paired to the inserted downstream sequence. pVE60hp (for VE60hp mRNA, see b in Fig. 1B) was derived from pS3hp. A 158-bp dsDNA was inserted to pS3hp between *Nde*I and *Bsr*GI sites to replace the existing hairpin completely.

pVE274hp (for VE274hp mRNA, see c in Fig. 1B) was derived from pVE60hp and constructed in several steps. Oligonucleotide A (60-base oligonucleotide), pGAAGAAGTG(GTN)$_{14}$GAAGAGGAG (p = phosphate, N = A, G, C or T; encoding Glu$_2$Val$_{15}$Glu$_3$), was annealed to an 18-base oligonucleotide connector, CACTTCTTCCTCCTCTTC, which was used to align multiple copies of oligonucleotide A in tandem, followed by ligation. To introduce primer sequences and functional sites, oligonucleotide B (CCATGGAACAAGCT-GAGGAATCAAACGTTCTTC) and oligonucleotide C (CGGCCGCCGCCC-TAGGGGCTGGCTTTCGCCAGCCCTTTAGTGAGGGGTTAATT) were ligated to the 5′ and 3′ sides, respectively. The ligated products were subjected to PCR (primers: TAGGAT*ccatgg*AACAAGCTGAGGAATCAAACG and AAGG-*CC*tgtacaAATTAACCCTCACTAAAGGGCTGG, *Nco*I and *Bsr*GI sites are shown in lower case, respectively) and then inserted to pVE60hp between *Nco*I and *Bsr*GI, replacing the existing hairpin sequence. The cloned sequences were dominated by two copies of oligonucleotide A. The oligonucleotide A regions from two clones were subjected to PCR and ligated to make a construct with essentially four copies of oligonucleotide A (called pRC4). To make a hairpin, the oligonucleotide A region of pRC4 was subjected to PCR (primers: GGACATCC-TAAT*tgtaca*AACAAGCTGAGGAATCAAACGTTCTTT and GCTTGTTA-*ggtctc*CCTAGGGAAACCTAGGGCGGCGGCCGCTCC, *Bsr*GI and *Bsa*I sites are shown in lower case, respectively; the latter creates a cohesive end compatible to *Avr*II), and inserted to pRC4 between *Avr*II and *Bsr*GI, downstream of the oligonucleotide A region. Note that the insertion is in an inverted (tail-to-tail) orientation, such that the insert is complementary to the oligonucleotide A region. RNA was synthesized *in vitro* with these plasmids with T7 RNA polymerase (Ambion). The synthesized RNA was purified by MEGAclear purification kit (Ambion) and stored at −30 °C.

Aminoacylation of tRNA. Purified tRNAs (tRNA^{Val}, tRNA^{Glu}, tRNA^{Phe} and tRNA^{fMet}) were purchased from Sigma. Total tRNA mixtures were purchased from Sigma and Roche. tRNAs were aminoacylated using DEAE-purified S-100 enzymes[38] and extracted with phenol/chloroform. The efficiency of aminoacylation for purified tRNAs was checked by acid gel electrophoresis[44]. Because of the heterogeneity, the efficiency of aminoacylation for the total tRNA mixtures cannot be determined from the gel.

Preparation of translation mixtures. The translation mixtures prepared here were used to inject into the reaction chamber to resume translation for the stalled ribosome–mRNA complexes. In general, the following components were mixed in a total of 45 μl of Buffer TL: 1 mM GTP, 5 mM PEP, 26.7 μM EF-Tu, and 0.02 mg ml^{-1} pyruvate kinase. The mixture was incubated at 37 °C for 15 min. Then 5 μl of Val-tRNA^{Val}, Glu-tRNA^{Glu} and Phe-tRNA^{Phe} (each with specified concentrations) were added, incubated at 37 °C for 5 min, and on ice for 5 min. The reaction was diluted with 350 μl (total 400 μl) Buffer TL containing 1 mM GTP, 1 mM ATP, 40 U RNAguard and 1 or 0.1 μM EF-G (final concentrations).

Finally, the mixture was filtered (0.22 μm, low protein binding, MILLEX-GV, Durapore membrane, Millipore) and put on ice.

Ribosome-releasing mixture. The ribosome-releasing mixture was used for control experiments to release the ribosome from the stalled ribosome–mRNA complex. The mixture contained the following components in 400 μl Buffer TL: 0.5 mM puromycin, 1 mM GTP, 0.25 μM EF-G, 1 μl DEAE-purified S-100 enzymes, and 40 U RNAguard.

Conversion of translation trajectories. To facilitate data analysis, the extension (end-to-end distance) changes of the VE274hp (or VE274hp-G2A) RNA during translation were converted to the number of bases translocated by the ribosome (base pairs broken), by an empirical method. The force–extension curves before translation were fitted to a worm-like chain (WLC) model[24] to obtain the basal WLC curve. A series of curves that mimicked the RNA elasticity corresponding to individual translation steps were calculated. This was done by adding six-nucleotide (three base pairs broken) components each time to the preceding WLC curves, starting from the basal WLC curve. This series of WLC curves was used to calculate the extension change for each step during translation with the consideration of gradual force decreasing. The theoretical relation between the number of bases translocated and extension change was well fitted to a double exponential equation. Finally, the fitted parameters were used to convert the experimentally measured extension change to the number of bases translocated. This empirical method worked well. For example, the force–extension curve after translation in Supplementary Fig. 3a (green trace) shows no apparent rips, suggesting that none or little residual hairpin was left. The converted translation trajectory (the blue trace in Fig. 3a) shows that the ribosome was stalled close to the apical loop of the hairpin, consistent with the experimental data.

Calculation of translation rates. The translation trajectories (number of bases translocated versus time; see Fig. 3a) were analysed by placing a moving window in the region corresponding to a translated Val patch. The window size was adjusted to cover 21 bp (seven codons, about half of a Val patch). The slope from linear regression for the data points in the window was recorded. The window was then moved to scan the whole region in the patch. The translation rate in this Val patch was finally assigned to the maximum slope. The same procedures were repeated for other translated Val patches on the trace.

Analysis of translocation steps. The 200 Hz unsmoothed extension trajectories obtained from the single-trap tweezers were analysed by placing a variable window around a translocation region corresponding to a single step. The window size was varied from 0.025 to 0.50 s, and the linear regression of the 5–100 points in the interval was calculated. The translocation time was assigned to the window size that best fitted a line with the maximum R^2 value.

36. Smith, S. B., Cui, Y. & Bustamante, C. Optical-trap force transducer that operates by direct measurement of light momentum. *Methods Enzymol.* **361**, 134–162 (2003).

37. Wen, J.-D. *et al.* Force unfolding kinetics of RNA using optical tweezers. I. Effects of experimental variables on measured results. *Biophys. J.* **92**, 2996–3009 (2007).

38. Moazed, D. & Noller, H. F. Interaction of tRNA with 23S rRNA in the ribosomal A, P, and E sites. *Cell* **57**, 585–597 (1989).

39. Lancaster, L. & Noller, H. F. Involvement of 16S rRNA nucleotides G1338 and A1339 in discrimination of initiator tRNA. *Mol. Cell* **20**, 623–632 (2005).

40. Wilson, K. S. & Noller, H. F. Mapping the position of translational elongation factor EF-G in the ribosome by directed hydroxyl radical probing. *Cell* **92**, 131–139 (1998).

41. Boon, K. *et al.* Isolation and functional analysis of histidine-tagged elongation factor Tu. *Eur. J. Biochem.* **210**, 177–183 (1992).

42. Traub, P., Mizushima, S., Lowry, C. V. & Nomura, M. in *RNA and Protein Synthesis* (ed. Moldave, K.) 521–539 (Academic Press, New York, 1981).

43. Culver, G. M. & Noller, H. F. *In vitro* reconstitution of 30S ribosomal subunits using complete set of recombinant proteins. *Methods Enzymol.* **318**, 446–460 (2000).

44. Varshney, U., Lee, C. P. & RajBhandary, U. L. Direct analysis of aminoacylation levels of tRNAs *in vivo*. Application to studying recognition of *Escherichia coli* initiator tRNA mutants by glutaminyl-tRNA synthetase. *J. Biol. Chem.* **266**, 24712–24718 (1991).

ARTICLES

A nuclear receptor-like pathway regulating multidrug resistance in fungi

Jitendra K. Thakur[1,2]*, Haribabu Arthanari[3]*, Fajun Yang[1,2]*, Shih-Jung Pan[4], Xiaochun Fan[3], Julia Breger[5], Dominique P. Frueh[3], Kailash Gulshan[6], Darrick K. Li[1], Eleftherios Mylonakis[5], Kevin Struhl[3], W. Scott Moye-Rowley[6], Brendan P. Cormack[4], Gerhard Wagner[3] & Anders M. Näär[1,2]

Multidrug resistance (MDR) is a serious complication during treatment of opportunistic fungal infections that frequently afflict immunocompromised individuals, such as transplant recipients and cancer patients undergoing cytotoxic chemotherapy. Improved knowledge of the molecular pathways controlling MDR in pathogenic fungi should facilitate the development of novel therapies to combat these intransigent infections. MDR is often caused by upregulation of drug efflux pumps by members of the fungal zinc-cluster transcription-factor family (for example Pdr1p orthologues). However, the molecular mechanisms are poorly understood. Here we show that Pdr1p family members in *Saccharomyces cerevisiae* and the human pathogen *Candida glabrata* directly bind to structurally diverse drugs and xenobiotics, resulting in stimulated expression of drug efflux pumps and induction of MDR. Notably, this is mechanistically similar to regulation of MDR in vertebrates by the PXR nuclear receptor, revealing an unexpected functional analogy of fungal and metazoan regulators of MDR. We have also uncovered a critical and specific role of the Gal11p/MED15 subunit of the Mediator co-activator and its activator-targeted KIX domain in antifungal/xenobiotic-dependent regulation of MDR. This detailed mechanistic understanding of a fungal nuclear receptor-like gene regulatory pathway provides novel therapeutic targets for the treatment of multidrug-resistant fungal infections.

Pathogenic fungi, especially *Candida* species, have emerged as important and prevalent opportunistic infections in individuals with compromised immunity, including those suffering from AIDS, cancer patients treated with chemotherapy, transplant recipients on immunosuppressive drugs and patients with advanced diabetes[1,2]. *Candida* species now account for 8–9% of all blood stream infections, with crude mortality rates of 40%[3,4]. Significantly, both intrinsic and acquired MDR is an important complication of fungal infections[5]. *C. glabrata*, which exhibits strong MDR, is emerging as a clinically important fungal pathogen, accounting, for example, for 20–24% of *Candida* blood stream infections in the USA[1,3]. There is thus an urgent need to elucidate the mechanisms underpinning MDR in pathogenic fungi to develop novel antifungal treatments. MDR in fungi is caused by the overexpression of membrane-spanning efflux pumps (for example the ATP-binding cassette (ABC) family of transporters), resulting in the expulsion of various structurally unrelated molecules[6]. Studies in *S. cerevisiae* have shown that the zinc-cluster transcription factor Pdr1p and the paralogue Pdr3p together confer resistance to several drugs and toxins through transcriptional activation of ABC transporter genes and members of the major facilitator superfamily of drug efflux pumps, including Pdr5p, Snq2p and Yor1p, as well as phospholipid-transfer proteins such as Pdr16p[7–15]. Several studies recently reported increased expression of drug efflux pumps in yeast in response to different xenobiotics, and demonstrated a requirement for Pdr1p/Pdr3p in this response[16–20], although the xenobiotic signalling pathway has not been elucidated. *C. glabrata* is evolutionarily closely related to *S. cerevisiae*, and recent studies have identified a Pdr1p orthologue (CgPdr1p) in *C. glabrata* that

regulates drug efflux pumps and controls MDR in this pathogen[21–23]. Increased knowledge of the mechanistic basis of CgPdr1p function in regulating multidrug resistance in *C. glabrata* could allow the identification of new classes of drugs to combat MDR in these clinically challenging infections.

Interestingly, the mammalian nuclear receptor pregnane X receptor (PXR) upregulates transcription of ABC transporters as well as cytochrome P450 detoxification enzymes in response to direct binding to structurally diverse xenobiotics[24,25]. However, similar direct transcription signalling mechanisms to regulate MDR have not been documented in non-vertebrate eukaryotes.

Transcription activators recruit co-activators that facilitate gene activation[26]. The Mediator co-activator, first characterized in yeast, interacts with RNA polymerase II and is involved in many transcriptional regulatory pathways[27]. The co-activator requirements, including a possible role of the Mediator, for xenobiotic-dependent transactivation of the MDR programme by Pdr1p orthologues have not been determined. Identification of co-activator targets for Pdr1p orthologues could facilitate the development of novel anti-MDR agents that target the activator/co-activator interface.

We show here that the Pdr1p orthologues in *S. cerevisiae* and *C. glabrata* directly bind xenobiotics to activate genes encoding drug efflux pumps, and exhibit functional characteristics that are mechanistically similar to the vertebrate xenobiotic receptor PXR. Moreover, our results demonstrate an essential and specific role for the Mediator co-activator subunit Gal11p (also known as MED15) in xenobiotic-dependent gene activation and MDR in *S. cerevisiae* and *C. glabrata*. The activation domains of Pdr1p orthologues bind

---

[1]Massachusetts General Hospital Cancer Center, Charlestown, Massachusetts 02129, USA. [2]Department of Cell Biology, [3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Department of Molecular Biology and Genetics, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA. [5]Division of Infectious Diseases, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [6]Department of Molecular Physiology and Biophysics, University of Iowa, Iowa City, Iowa 52242, USA.
*These authors contributed equally to this work.

directly to a domain present in Gal11p that is structurally conserved with the activator-binding KIX domain found in the human ARC105/MED15 Mediator subunit and in vertebrate CBP/p300 acetyltransferases. Our results demonstrate that fungi harbour a nuclear receptor-like pathway controlling MDR, which could represent a novel therapeutic target for the treatment of MDR in pathogenic fungi such as *C. glabrata*.

Pdr1p and Pdr3p are xenobiotic receptors

The expression of ATP-dependent drug efflux pumps (for example *PDR5*) and other Pdr1p/Pdr3p target genes (for example *PDR16*) in *S. cerevisiae* can be induced by chemically distinct drugs and xenobiotics, including the antifungal ketoconazole, the translation inhibitor cycloheximide and the classic PXR agonist rifampicin, in a Pdr1p/Pdr3p-dependent manner (Fig. 1a and data not shown). In contrast, the glucocorticoid receptor agonist dexamethasone was consistently a poor inducer of Pdr1p/Pdr3p-regulated genes (Fig. 1a). Because the mammalian nuclear receptor PXR controls MDR gene expression by direct binding to xenobiotics[24,25], and based on the intriguing functional similarities of Pdr1p/Pdr3p to PXR, we tested whether Pdr1p and Pdr3p could also directly interact with xenobiotics to stimulate expression of their target genes. Importantly, immunopurified Pdr1p binds ketoconazole with a dissociation constant (K_D) of about 39 μM, similar to the range of binding affinities (mid-nanomolar to high micromolar) of ligands for mammalian PXR (Fig. 1b)[24]. Binding of radiolabelled ketoconazole to Pdr1p was effectively competed by unlabelled xenobiotics that activate Pdr1p/Pdr3p target genes *in vivo*, including rifampicin,

cycloheximide and ketoconazole itself (Fig. 1b, c). Similar results were obtained with the Pdr1p paralogue Pdr3p, consistent with its functional redundancy in gene expression assays with these xenobiotics (Supplementary Fig. 2a, b and data not shown). Deletion analysis revealed that a small region in Pdr1p (amino acids 352–543) carboxy (C)-terminal to the DNA-binding domain is sufficient for binding to ketoconazole (Fig. 1d and Supplementary Fig. 2c, d). A similar region in Pdr3p also mediates ketoconazole binding, supporting the notion of a discrete xenobiotic-binding domain (XBD) in this family of zinc-cluster transcription factors (Supplementary Fig. 2e, f). Nuclear receptors harbour autonomous ligand-binding domains and ligand-responsive activation domains located C-terminal to the zinc finger DNA-binding domain that can be transferred to heterologous DNA-binding domains. Fusion experiments to the yeast Gal4p DNA-binding domain (Gal4pDBD) identified a large C-terminal domain in Pdr1p (amino acids 352–1063), encompassing both the XBD and the C-terminal activation domain, as the minimal transferable xenobiotic-responsive transactivation domain (Fig. 1d and Supplementary Fig. 3a, b). These results demonstrate that yeast Pdr1p/Pdr3p activate transcription of target genes in response to direct binding to specific xenobiotics by a discrete transferable ligand-binding domain, which suggests that these critical transcription regulators of MDR in yeast function in a manner analogous to the vertebrate nuclear receptor PXR.

Requirement for Gal11p in xenobiotic response and MDR

Next, we studied the transcriptional co-activator requirements for ligand-dependent gene activation by Pdr1p/Pdr3p. The Mediator co-activator plays critical roles in transcriptional activation, from yeast to human[27]. We therefore examined whether the Mediator is involved in Pdr1p/Pdr3p-dependent and xenobiotic-stimulated gene activation and MDR. Although deletion of most Mediator subunits caused few or modest effects on MDR, deletion of the gene encoding the Gal11p subunit resulted in striking sensitivity to several toxins/xenobiotics, including ketoconazole, cycloheximide and 4-nitroquinoline oxide (Fig. 2a and data not shown). Consistent with these findings, xenobiotic-dependent expression of the *PDR5* and *PDR16* genes was specifically and strongly decreased in the *GAL11* deletion strain, similar to that observed with the *PDR1/3* deletion strain (Fig. 2b and data not shown). Moreover, deletion of the *GAL11* gene caused strongly decreased activation of the Pdr1p/Pdr3p-responsive promoters from the *PDR5* and *SNQ2* genes by constitutively active Pdr1p and Pdr3p mutants isolated from multidrug-resistant yeast (Supplementary Fig. 4)[9,28,29]. Co-immunoprecipitation experiments showed that Myc$_6$–Pdr1p interacts with Gal11p–Flag$_2$ in a xenobiotic-stimulated manner *in vivo* (Fig. 2c), and chromatin immunoprecipitation data indicate that Gal11p is specifically recruited to Pdr1p/Pdr3p target genes in a Pdr1p/Pdr3p-dependent fashion (Supplementary Fig. 5 and data not shown). These studies demonstrate that Gal11p is essential for xenobiotic-dependent gene activation and MDR mediated by Pdr1p/Pdr3p.

Gal11p harbours an activator-binding KIX domain

Gal11p orthologues contain sequences in their amino termini that exhibit significant homology to the activator-binding KIX domain of the metazoan Mediator subunit ARC105/MED15 (Supplementary Fig. 6a)[30,31]. The KIX domain was first identified as an activator target in the histone acetyltransferases CBP/p300, mediating interactions with many metazoan transcription factors[32]. Based on the documented functional importance of the KIX domains in ARC105/MED15 and CBP/p300 in specific gene-activation pathways[31,33–35], we attempted to identify activators in yeast that interact with the predicted Gal11p KIX domain. Remarkably, affinity chromatography of yeast whole-cell extract over the putative Gal11p KIX domain yielded a single specific band that was identified by mass spectrometry as Pdr1p (Supplementary Fig. 6b, c), demonstrating that Pdr1p interacts strongly with the putative Gal11p KIX domain.
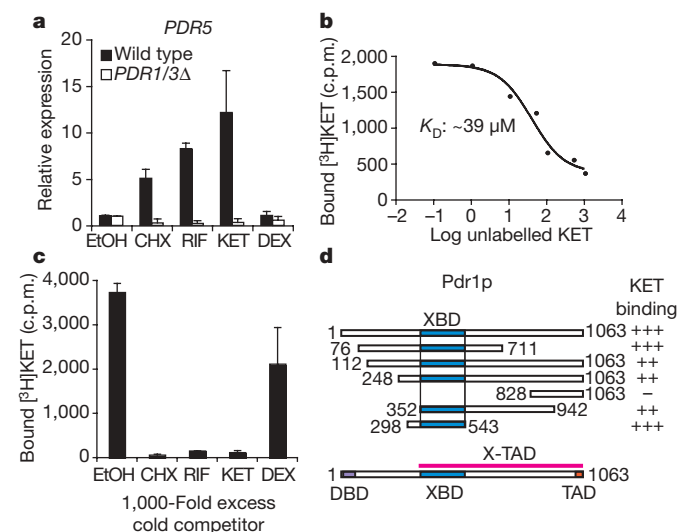
Figure 1 | Pdr1p is a xenobiotic receptor. a, Xenobiotic-induced transcription of the *PDR5* gene is dependent on Pdr1p/Pdr3p. Quantitative real-time RT–PCR reactions were performed in triplicate. Mean values from three independent experiments are shown; error bars, s.d. EtOH, ethanol, vehicle; CHX, cycloheximide; RIF, rifampicin; KET, ketoconazole; DEX, dexamethasone. **b**, Pdr1p directly binds to the antifungal ketoconazole. Cold ligand competition was used to determine the binding affinity of [³H]KET to Myc$_6$–Pdr1p. [³H]KET (0.1 μM) was used for this experiment. Logarithmic concentration of unlabelled KET is displayed on the *x* axis. Mean values from triplicate samples are shown. The graph was generated using GraphPad Prism4. **c**, Cold competition assay reveals binding of various xenobiotics to Myc$_6$–Pdr1p. Unlabelled xenobiotics were used for competition (*x* axis). The experiment was performed in triplicate. Mean values are shown; error bars, s.d. **d**, The xenobiotic-binding domain of Pdr1p. Glutathione–Sepharose beads with GST–Pdr1p protein fragments were used for direct binding assay using 0.1 μM of [³H]KET. Symbols: +, binding; −, no binding. Lower cartoon shows positions of the DNA-binding domain (DBD), the xenobiotic-binding domain (XBD), the transactivation domain (TAD) and the xenobiotic-responsive transactivation domain (X-TAD).

We have also found that the putative Gal11p KIX binds to purified Pdr1p and that this interaction is further enhanced by xenobiotics (Fig. 2d).
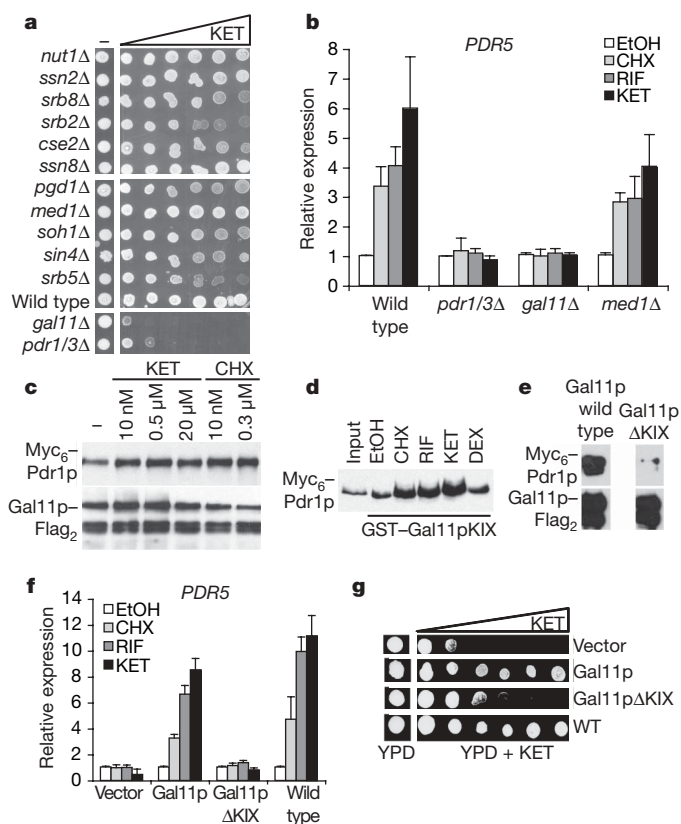
Figure 2 | Requirement for the Gal11p Mediator subunit and its KIX domain in Pdr1p/Pdr3p-dependent transcription of target genes and MDR.
a, Growth of yeast cells on 1% yeast extract, 2% peptone and 2% dextrose (YPD) with increasing concentrations of ketoconazole reveals a specific requirement for Gal11p for ketoconazole resistance; −, yeasts growing on YPD without ketoconazole. **b**, Gal11p is required for Pdr1p/Pdr3p-mediated and xenobiotic-dependent transcription of the *PDR5* gene. Deletion of *PDR1/3* was used as positive control. Real-time quantitative RT–PCR reactions were performed in triplicate. Mean values are shown; error bars, s.d. **c**, Co-immunoprecipitation shows the interaction between Myc6–Pdr1p and Gal11p–Flag2 in the absence and presence of different concentrations of ketoconazole and cycloheximide. Anti-Flag antibodies were used for immunoprecipitation, with IgG as negative control (no significant binding: data not shown). Anti-Myc (upper panel) and anti-Flag (lower panel) antibodies were used for immunodetection. There is no significant difference in the input material (not shown). **d**, GST-pulldown analysis demonstrates increased interaction between the GST–Gal11pKIX domain and Myc6–Pdr1p in the presence of activating xenobiotics. Bound Myc6–Pdr1p was detected by anti-Myc immunoblotting. **e**, Co-immunoprecipitation shows the importance of the Gal11p KIX domain for *in vivo* interaction between Gal11p and Pdr1p. Yeast extracts from cells cultured in the presence of ketoconazole expressing Myc6–Pdr1p and either Gal11p–Flag2 or Gal11pΔKIX–Flag2 were used for co-immunoprecipitation. Immunoprecipitation used anti-Flag antibodies; immunodetection used anti-Myc (upper panel) or anti-Flag (lower panel) antibodies. **f**, The Gal11p KIX domain is required for xenobiotic-induced transcription of *PDR5*. The *gal11Δ* yeast strain was reconstituted with plasmids expressing full-length Gal11p, or Gal11p lacking the KIX domain (amino acids 1–100), or with vector control. Wild-type yeast was used as positive control. Real-time quantitative RT–PCR reactions were performed in triplicate. Mean values are shown; error bars, s.d. **g**, Growth of yeast cells on YPD with increasing concentrations of ketoconazole shows the requirement of the Gal11p KIX domain for ketoconazole resistance. The *gal11Δ* yeast strain was transformed with plasmids harbouring either full-length Gal11p, Gal11p lacking the KIX domain (amino acids 1–100) or vector control. Wild-type yeast was used as positive control. Left panel shows yeast cells grown on YPD without ketoconazole.

Mapping studies revealed that C-terminal Pdr1p sequences containing the activation domain (Pdr1pAD) bind to the Gal11p KIX domain (Supplementary Fig. 6d and data not shown). Interestingly, Pdr1pAD also bound to the CBP and ARC105/MED15 KIX domains (Supplementary Fig. 6d). Consistent with its ability to engage mammalian co-activators, Pdr1pAD fused to Gal4pDBD mediated potent gene activation in human cells (Supplementary Fig. 6e). The Gal11p KIX domain can also interact with the human SREBP-1a activator that we previously showed associates with the ARC105/MED15 and CBP KIX domains (Supplementary Fig. 6b, d)[31,36]. In contrast, the CBP/p300 KIX-binding activators CREB and c-Myb cannot interact with the Gal11p KIX domain, nor with the human ARC105/MED15 KIX domain (Supplementary Fig. 4d)[31,37,38]. These results indicate that the putative Gal11p KIX domain is a specific target only for certain activators, and functionally behaves more like the human ARC105/MED15 KIX domain than the CBP KIX domain, in keeping with the fact that both ARC105/MED15 and Gal11p are components of the Mediator family of co-activators. Our studies also revealed the functional importance of the Gal11p KIX domain for Pdr1p/Pdr3p gene activation and MDR *in vivo*. Deletion of the Gal11p KIX domain strongly decreased interaction of Gal11p with Pdr1p in co-immunoprecipitation experiments (Fig. 2e). Moreover, exogenous expression of wild-type Gal11p, but not KIX-deleted Gal11p, can rescue both xenobiotic-dependent activation of Pdr1p/Pdr3p target genes and resistance to ketoconazole in yeast deleted for *GAL11* (Fig. 2f, g and data not shown).

To provide molecular details that could yield further insights into the gene activation mechanism by Pdr1p, we have determined the solution structure of the Gal11p Pdr1p-binding domain by NMR. The high-resolution structure reveals a three-helix bundle fold with marked similarity to the human ARC105/MED15 and mouse CBP KIX domains (Fig. 3a, b)[31,35]. Like the mammalian KIX domains, the three helices in the Gal11p KIX domain pack an extensively hydrophobic core (Supplementary Fig. 7c). In the ARC105/MED15 and CBP KIX structures, hydrophobic patches on the surface of the KIX domains mediate interactions with several activators (Supplementary Fig. 9)[31,35,38]. We show here by chemical shift analysis that peptides containing the C-terminal 12 and 34 amino acids of the Pdr1pAD also interact with amino acids within a large hydrophobic groove contributed by all three helices, including L25, Q26, M29, I31, L34, A42, I47, N51, F52 A55, V74, A75 and V76 (Fig. 3c and Supplementary Figs 7–10). Binding studies of Pdr1p with point-mutated Gal11p KIX proteins in the presence of ketoconazole also revealed several KIX amino acids as being important for Pdr1p binding, consistent with the NMR data (Fig. 3c–e). NMR showed that yeast Pdr1pAD-34 can also interact with the human ARC105/MED15 KIX domain, consistent with the binding of a larger Pdr1pAD fragment (Supplementary Figs 6d and 10a). Analysis by chemical shift perturbation revealed that the Pdr1p activation domain interaction surface on the Gal11p KIX domain substantially overlaps with that of the human SREBP-1a activation domain (Supplementary Fig. 10d). These results suggest similarities in the way activators target orthologous KIX domains. However, there are also significant differences in the way Pdr1p and SREBP-1a engage their cognate KIX domains (Supplementary Fig. 9). These results agree with earlier observations that different activation domains (for example CREB pKID, c-Myb and MLL) bind both overlapping and distinct epitopes on the human ARC105/MED15 and mouse CBP KIX domains (Supplementary Fig. 9)[31,35,38–43]. Thus, although the structure of the KIX domain is conserved between mammals and yeast, a variety of interfaces on KIX domains are used to accommodate various activation domains. Taken together, our findings reveal that the activator-binding domain in Gal11p indeed folds into a functionally conserved KIX domain; they also pinpoint key residues in the Gal11p KIX domain involved in binding to the Pdr1p activation domain.

Conservation of xenobiotic gene regulation in *C. glabrata*

Having dissected the molecular mechanisms underpinning the xenobiotic gene regulatory network controlling MDR in the non-pathogenic yeast *S. cerevisiae*, we next determined the potential clinical relevance of our findings for pathogenic fungi. *C. glabrata* is the second most common cause of invasive candidiasis, and has been reported to exhibit intrinsic MDR, in particular to azoles[3,44,45]. *C. glabrata* harbours a highly conserved Pdr1p orthologue that also regulates drug efflux pumps in response to xenobiotics[22,23]. Based on our results with Pdr1p/Pdr3p in *S. cerevisiae*, we hypothesized that CgPdr1p might also directly bind to azoles and other xenobiotics to promote gene expression and MDR in *C. glabrata*.

Expression of the *C. glabrata* drug efflux pump gene *CDR2* is stimulated by xenobiotics in a CgPdr1p-dependent manner, and CgPdr1p is required for the intrinsically high azole resistance of *C. glabrata* (Fig. 4a, b and Supplementary Fig. 11a)[22,23]. We performed complementation experiments in *PDR1/3*-deleted *S. cerevisiae* to test whether expression of CgPdr1p in this strain could functionally substitute for *S. cerevisiae* Pdr1p/Pdr3p. Indeed, CgPdr1p expression rescued both xenobiotic-dependent gene activation and MDR to a similar extent as the expression of ScPdr1p, establishing the functional similarity of these transcription factors (Fig. 4c and data not shown). Importantly, like its *S. cerevisiae* orthologues, CgPdr1p binds directly to radiolabelled ketoconazole, indicating that CgPdr1p also acts by a direct effector mechanism akin to nuclear receptor signalling (Fig. 4d).

Next, we examined the co-activator requirements for xenobiotic-dependent gene activation and MDR in *C. glabrata*. Interestingly, *C. glabrata* harbours two distinct genes with significant sequence

similarity to the *S. cerevisiae GAL11* gene (termed *CgGAL11A* and *CgGAL11B* here; Supplementary Fig. 11b). Deletion of the *CgGAL11A* gene strongly decreased xenobiotic-dependent activation of *CgCDR2*, similar to the effects of deleting *CgPDR1*, whereas deletion of the *CgGAL11B* gene had no effect on *CgCDR2* expression (Fig. 4a). *CgGAL11A* deletion also abrogated MDR, causing markedly increased sensitivity to azoles such as ketoconazole and fluconazole, as well as to cycloheximide, as revealed by growth assays (Fig. 4b and data not shown). The *CgGAL11A/B* double-deletion strain also performed like the *CgGAL11A* deletion strain (Fig. 4a, b); these data suggest that CgGal11Ap is functionally more important in CgPdr1p xenobiotic-dependent gene activation than CgGal11Bp. Consistent with this notion, a CgPdr1p activation domain fragment interacts better with the CgGal11Ap KIX domain than with the CgGal11Bp KIX domain (Supplementary Fig. 11c).

Non-mammalian hosts, such as the nematode *Caenorhabditis elegans*, have recently been shown to provide powerful and facile model systems to investigate fungal pathogenicity, MDR mechanisms, host response pathways and to identify novel antifungals[46,47]. Infection of *C. elegans* with *C. glabrata* and other *Candida* species results in the death of most nematodes within six days, even in the presence of the antifungal fluconazole (Fig. 4e)[46]. The deletion of *CgPDR1*, *CgGAL11A*, *CgGAL11B* or *CgMED1* had little effect on the pathogenicity of *C. glabrata* in this model organism in the absence of antifungals (data not shown). By contrast, in the presence of fluconazole, we observed that nematodes infected with the *Cgpdr1* and *Cggal11A* deletion strains exhibited significantly increased survival (Fig. 4e). *C. elegans* infected with the *Cggal11B* or *Cgmed1* deletion strains showed little difference in survival rates in the presence of
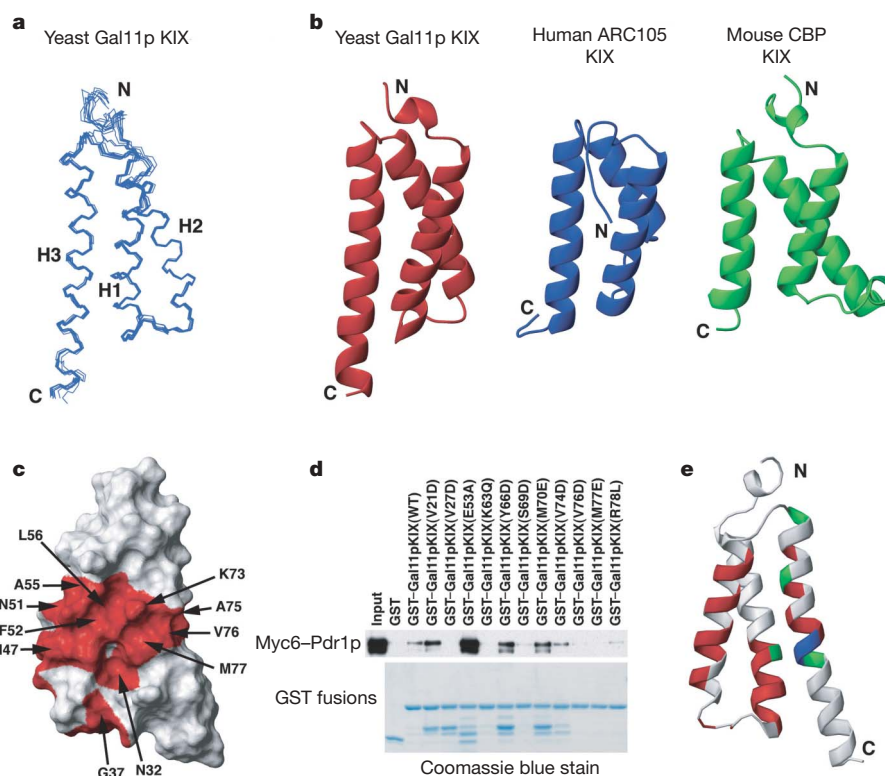
Figure 3 | Structure of the Gal11p KIX domain and structural/mutational analysis of the Pdr1pAD-Gal11p KIX interface. a, Representation showing the ten lowest energy solution structures of the Gal11p KIX domain. **b,** Ribbon diagram of the mean solution structures of yeast Gal11p, human ARC105/MED15 and mouse CBP KIX domains. **c,** Surface representation of Gal11p KIX with the Pdr1pAD-12 interaction surface (molar ratio 5:1 Gal11p KIX:Pdr1pAD-12) coloured red. The residues shown in red correspond to a chemical shift change of more than 0.02 p.p.m. **d,** Binding of Myc6–Pdr1p to point-mutated GST–Gal11p KIX domain fusion proteins in

a GST-pulldown assay. Top panel, Myc6–Pdr1p binding as detected by anti-Myc immunoblotting. Bottom panel, GST–Gal11p KIX fusion proteins as detected by Coomassie blue staining. **e,** Ribbon representation of Gal11p KIX where the residues shown in red correspond to a chemical shift change of more than 0.02 p.p.m. upon addition of Pdr1pAD-12 (molar ratio of 5:1 Gal11p KIX:Pdr1pAD-12). Residues whose mutation disrupts binding to Myc6–Pdr1p are shown in green; those in blue represent the residues implicated in Pdr1p binding by NMR and mutational studies.
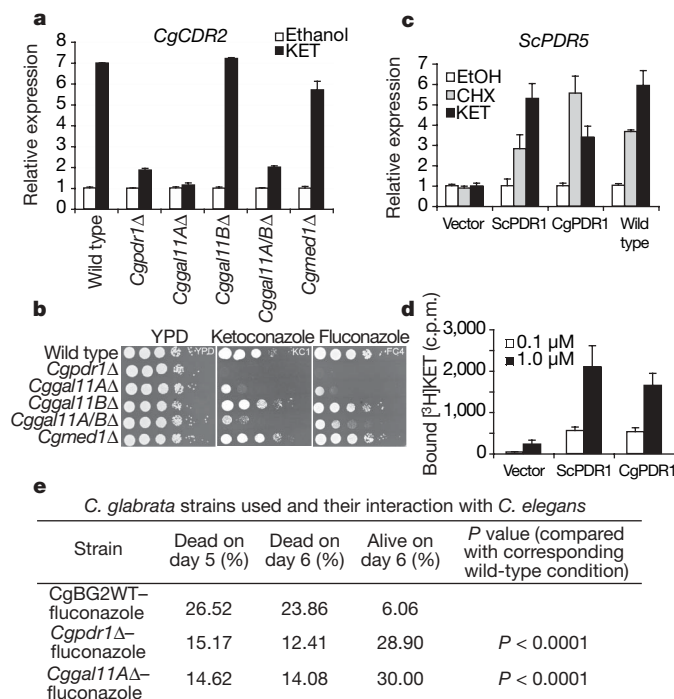
Figure 4 | Dissection of the molecular mechanism of drug resistance in *C. glabrata*. a, CgPdr1p and CgGal11Ap are required for ketoconazole-induced transcription of *CgCDR2*. Wild-type and *Cgmed1Δ* strains were used as controls. Real-time quantitative RT–PCR reactions were performed in triplicate. Mean values are shown; error bars, s.d. **b**, Growth of wild-type or mutant *C. glabrata* cells on YPD containing either ethanol vehicle (YPD), ketoconazole (1 μg ml^{-1}; KC1) or fluconazole (4 μg ml^{-1}; FC4) shows that CgPdr1p and CgGal11Ap are required for azole resistance in *C. glabrata*. **c**, CgPdr1p can functionally complement ScPdr1p/Pdr3p for drug-induced *PDR5* transcription. *S. cerevisiae pdr1/3Δ* double-deletion mutant strains transformed with plasmids harbouring either ScPDR1 or CgPDR1 complementary DNA (cDNA) or vector were used for this assay. Real-time quantitative RT–PCR reactions were performed in triplicate. Mean values are shown; error bars, s.d. **d**, Binding of [³H]KET with CgPdr1p. Beads with immunopurified Myc$_6$–CgPdr1p or Myc$_6$–ScPdr1p were used for binding assays. Mean values from triplicate experiments are shown; error bars, s.d. **e**, Effect of fluconazole on killing of *C. elegans* by wild-type or mutant strains of *C. glabrata* on days 5 and 6 (120 and 144 h). Fluconazole increased the lifespan of nematodes when the *CgPDR1* and *CgGAL11A* genes were deleted in *C. glabrata*. *P* values were calculated based on the entire 6-day experiment, with log-rank and Wilcoxon tests performed by STATA 6 statistical software (Stata).

fluconazole, similar to wild-type *C. glabrata* (data not shown). These results are consistent with the *in vitro* findings and demonstrate a critical role for the functional interaction of CgPdr1p and CgGal11Ap in *C. glabrata* MDR *in vivo* in a fungal pathogenesis model.

Discussion

We have found that fungi harbour sensor/effector regulatory mechanisms governing detoxification response that exhibit intriguing functional similarities to vertebrate nuclear xenobiotic receptors. Pdr1p orthologues and PXR both bind directly to structurally unrelated xenobiotics and drugs. As a result, they activate the expression of genes encoding ATP-dependent drug efflux pumps (for example P-glycoprotein/MDR1 orthologues) (Supplementary Fig. 1)[24]. Bioinformatics studies based on the conservation of the zinc finger DNA-binding domain and the ligand-binding domain have shown that nuclear receptors first arose during metazoan evolution[48]. Our results revealing functional similarities of fungal Pdr1p orthologues and vertebrate PXR do not constitute proof of evolutionary orthology. However, taken together with a recent study showing that

the yeast zinc-cluster family member Oaf1p may function similarly to the vertebrate PPARα nuclear receptor[49], they suggest that further studies of mechanistic analogies (and possible evolutionary relationships) between fungal zinc-cluster transcription factors and metazoan nuclear receptors are warranted.

We have shown here that fungal Pdr1p orthologues interact physically and functionally with the Gal11p/MED15 subunit of the Mediator. Our data revealing that Gal11p harbours an activator-binding domain with marked structural similarity to KIX domains present in metazoan co-activators, indicate strong evolutionary conservation, which implies critical functionality. Interestingly, a recent study showed that the *C. elegans* nuclear receptor NHR-49 interacts with the KIX-containing Gal11p orthologue MDT-15 and requires MDT-15 for activation of fatty-acid metabolism genes[50]. This raises the question of whether other metazoan nuclear receptors also use the ARC105/MED15 subunit as a transducer of gene-activating signals. The targeting of ARC105/MED15 orthologues in fungi and metazoans might thus represent an ancient mechanism of activation by ligand-dependent transcription factors.

The elucidation of the molecular mechanism of xenobiotic-dependent regulation of MDR by CgPdr1p in *C. glabrata* should provide novel targets for the development of 'co-therapeutics' that augment standard antifungal therapies by directly interfering with the mechanistic underpinnings of antifungal-induced MDR (Supplementary Fig. 1). For example, small-molecule antagonists might be identified that lock Pdr1p orthologues in an inactive conformation, thereby preventing activation of the efflux pump genes and resulting in sensitization to standard antifungal therapy. Alternatively, the highly hydrophobic groove in the Gal11p KIX domain that provides the Pdr1p AD docking site might serve as a promising therapeutic target. Our findings should also provide the foundation for studies investigating whether similar regulatory mechanisms govern MDR in other clinically significant pathogenic fungi, such as *Candida albicans*.

METHODS SUMMARY

Details of plasmids, primers, fungal strains, xenobiotic treatment, quantitative real-time polymerase chain reaction with reverse transcription (qRT–PCR), β-galactosidase assay, drug sensitivity assays, immunoprecipitation and immunoblotting, chromatin immunoprecipitation, drug binding and cold competition assays, glutathione *S*-transferase (GST) fusion-protein generation and pulldown assays, *C. elegans* fungal killing assays and structural analysis by NMR are given in Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Richardson, M. D. Changing patterns and trends in systemic fungal infections. *J. Antimicrob. Chemother.* **56**, i5–i11 (2005).
2. Aperis, G., Myriounis, N., Spanakis, E. K. & Mylonakis, E. Developments in the treatment of candidiasis: more choices and new challenges. *Expert Opin. Investig. Drugs* **15**, 1319–1336 (2006).
3. Pfaller, M. A. & Diekema, D. J. Epidemiology of invasive candidiasis: a persistent public health problem. *Clin. Microbiol. Rev.* **20**, 133–163 (2007).
4. Wisplinghoff, H. *et al.* Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clin. Infect. Dis.* **39**, 309–317 (2004).
5. Klepser, M. E. *Candida* resistance and its clinical relevance. *Pharmacotherapy* **26**, 68S–75S (2006).
6. Sipos, G. & Kuchler, K. Fungal ATP-binding cassette (ABC) transporters in drug resistance & detoxification. *Curr. Drug Targets* **7**, 471–481 (2006).
7. Balzi, E. *et al.* The multidrug resistance gene PDR1 from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **262**, 16871–16879 (1987).
8. Meyers, S. *et al.* Interaction of the yeast pleiotropic drug resistance genes PDR1 and PDR5. *Curr. Genet.* **21**, 431–436 (1992).
9. Balzi, E. *et al.* PDR5, a novel yeast multidrug resistance conferring transporter controlled by the transcription regulator PDR1. *J. Biol. Chem.* **269**, 2206–2214 (1994).
10. Katzmann, D. J. *et al.* Transcriptional control of the yeast PDR5 gene by the PDR3 gene product. *Mol. Cell. Biol.* **14**, 4653–4661 (1994).

11. Delaveau, T. *et al.* PDR3, a new yeast regulatory gene, is homologous to PDR1 and controls the multidrug resistance phenomenon. *Mol. Gen. Genet.* **244**, 501–511 (1994).

12. Decottignies, A. *et al.* Identification and characterization of SNQ2, a new multidrug ATP binding cassette transporter of the yeast plasma membrane. *J. Biol. Chem.* **270**, 18150–18157 (1995).

13. Katzmann, D. J. *et al.* Expression of an ATP-binding cassette transporter-encoding gene (YOR1) is required for oligomycin resistance in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **15**, 6875–6883 (1995).

14. van den Hazel, H. B. *et al.* PDR16 and PDR17, two homologous genes of *Saccharomyces cerevisiae*, affect lipid biosynthesis and resistance to multiple drugs. *J. Biol. Chem.* **274**, 1934–1941 (1999).

15. Moye-Rowley, W. S. Transcriptional control of multidrug resistance in the yeast *Saccharomyces*. *Prog. Nucleic Acid Res. Mol. Biol.* **73**, 251–279 (2003).

16. Mamnun, Y. M., Schuller, C. & Kuchler, K. Expression regulation of the yeast PDR5 ATP-binding cassette (ABC) transporter suggests a role in cellular detoxification during the exponential growth phase. *FEBS Lett.* **559**, 111–117 (2004).

17. Gao, C., Wang, L., Milgrom, E. & Shen, W. C. On the mechanism of constitutive Pdr1 activator-mediated PDR5 transcription in *Saccharomyces cerevisiae*: evidence for enhanced recruitment of coactivators and altered nucleosome structures. *J. Biol. Chem.* **279**, 42677–42686 (2004).

18. Lucau-Danila, A. *et al.* Early expression of yeast genes affected by chemical stress. *Mol. Cell. Biol.* **25**, 1860–1868 (2005).

19. Alenquer, M., Tenreiro, S. & Sa-Correia, I. Adaptive response to the antimalarial drug artesunate in yeast involves Pdr1p/Pdr3p-mediated transcriptional activation of the resistance determinants TPO1 and PDR5. *FEMS Yeast Res.* **6**, 1130–1139 (2006).

20. Fardeau, V. *et al.* The central role of PDR1 in the foundation of yeast drug resistance. *J. Biol. Chem.* **282**, 5063–5074 (2007).

21. Vermitsky, J. P. & Edlind, T. D. Azole resistance in *Candida glabrata*: coordinate upregulation of multidrug transporters and evidence for a Pdr1-like transcription factor. *Antimicrob. Agents Chemother.* **48**, 3773–3781 (2004).

22. Tsai, H. F., Krol, A. A., Sarti, K. E. & Bennett, J. E. *Candida glabrata* PDR1, a transcriptional regulator of a pleiotropic drug resistance network, mediates azole resistance in clinical isolates and petite mutants. *Antimicrob. Agents Chemother.* **50**, 1384–1392 (2006).

23. Vermitsky, J. P. *et al.* Pdr1 regulates multidrug resistance in *Candida glabrata*: gene disruption and genome-wide expression studies. *Mol. Microbiol.* **61**, 704–722 (2006).

24. Kliewer, S. A., Goodwin, B. & Willson, T. M. The nuclear pregnane X receptor: a key regulator of xenobiotic metabolism. *Endocr. Rev.* **23**, 687–702 (2002).

25. Willson, T. M. & Kliewer, S. A. PXR, CAR and drug metabolism. *Nature Rev. Drug Discov.* **1**, 259–266 (2002).

26. Näär, A. M., Lemon, B. D. & Tjian, R. Transcriptional coactivator complexes. *Annu. Rev. Biochem.* **70**, 475–501 (2001).

27. Kornberg, R. D. Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.* **30**, 235–239 (2005).

28. Kean, L. S. *et al.* Plasma membrane translocation of fluorescent-labeled phosphatidylethanolamine is controlled by transcription regulators, PDR1 and PDR3. *J. Cell Biol.* **138**, 255–270 (1997).

29. Gulshan, K., Rovinsky, S. A., Coleman, S. T. & Moye-Rowley, W. S. Oxidant-specific folding of Yap1p regulates both transcriptional activation and nuclear localization. *J. Biol. Chem.* **280**, 40524–40533 (2005).

30. Novatchkova, M. & Eisenhaber, F. Linking transcriptional mediators via the GACKIX domain super family. *Curr. Biol.* **14**, R54–R55 (2004).

31. Yang, F. *et al.* An ARC/Mediator subunit required for SREBP control of cholesterol and lipid homeostasis. *Nature* **442**, 700–704 (2006).

32. Goodman, R. H. & Smolik, S. CBP/p300 in cell growth, transformation, and development. *Genes Dev.* **14**, 1553–1577 (2000).

33. Kasper, L. H. *et al.* A transcription-factor-binding surface of coactivator p300 is required for haematopoiesis. *Nature* **419**, 738–743 (2002).

34. Kasper, L. H. *et al.* Conditional knockout mice reveal distinct functions for the global transcriptional coactivators CBP and p300 in T-cell development. *Mol. Cell. Biol.* **26**, 789–809 (2006).

35. Radhakrishnan, I. *et al.* Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell* **91**, 741–752 (1997).

36. Näär, A. M. *et al.* Chromatin, TAFs, and a novel multiprotein coactivator are required for synergistic activation by Sp1 and SREBP-1a in vitro. *Genes Dev.* **12**, 3020–3031 (1998).

37. Dai, P. *et al.* CBP as a transcriptional coactivator of c-Myb. *Genes Dev.* **10**, 528–540 (1996).

38. Zor, T., De Guzman, R. N., Dyson, H. J. & Wright, P. E. Solution structure of the KIX domain of CBP bound to the transactivation domain of c-Myb. *J. Mol. Biol.* **337**, 521–534 (2004).

39. De Guzman, R. N., Goto, N. K., Dyson, H. J. & Wright, P. E. Structural basis for cooperative transcription factor binding to the CBP coactivator. *J. Mol. Biol.* **355**, 1005–1013 (2006).

40. Goto, N. K. *et al.* Cooperativity in transcription factor binding to the coactivator CREB-binding protein (CBP). The mixed lineage leukemia protein (MLL) activation domain binds to an allosteric site on the KIX domain. *J. Biol. Chem.* **277**, 43168–43174 (2002).

41. Parker, D. *et al.* Analysis of an activator:coactivator complex reveals an essential role for secondary structure in transcriptional activation. *Mol. Cell* **2**, 353–359 (1998).

42. Parker, D. *et al.* Role of secondary structure in discrimination between constitutive and inducible activators. *Mol. Cell. Biol.* **19**, 5601–5607 (1999).

43. Radhakrishnan, I. *et al.* Structural analyses of CREB–CBP transcriptional activator–coactivator complexes by NMR spectroscopy: implications for mapping the boundaries of structural domains. *J. Mol. Biol.* **287**, 859–865 (1999).

44. Prasad, R., Gaur, N. A., Gaur, M. & Komath, S. S. Efflux pumps in drug resistance of *Candida*. *Infect. Disord. Drug Targets* **6**, 69–83 (2006).

45. Pfaller, M. A. *et al.* Results from the ARTEMIS DISK Global Antifungal Surveillance study, 1997 to 2005: an 8.5-year analysis of susceptibilities of *Candida* species and other yeast species to fluconazole and voriconazole determined by CLSI standardized disk diffusion testing. *J. Clin. Microbiol.* **45**, 1735–1745 (2007).

46. Breger, J. *et al.* Antifungal chemical compounds identified using a *C. elegans* pathogenicity assay. *PLoS Pathogens* **3**, e18 (2007).

47. Mylonakis, E. & Aballay, A. Worms and flies as genetically tractable animal models to study host–pathogen interactions. *Infect. Immun.* **73**, 3833–3841 (2005).

48. Bertrand, S. *et al.* Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. *Mol. Biol. Evol.* **21**, 1923–1937 (2004).

49. Phelps, C. *et al.* Fungi and animals may share a common ancestor to nuclear receptors. *Proc. Natl Acad. Sci. USA* **103**, 7077–7081 (2006).

50. Taubert, S., Van Gilst, M. R., Hansen, M. & Yamamoto, K. R. A Mediator subunit, MDT-15, integrates regulation of fatty acid metabolism by NHR-49-dependent and -independent pathways in *C. elegans*. *Genes Dev.* **20**, 1137–1149 (2006).

METHODS

Media and chemicals. All bacteria were routinely grown in Luria-Bertani broth medium with required antibiotics. *S. cerevisiae* cells were grown either in YPD or in synthetic defined medium as required. For culturing *C. glabrata*, we used synthetic complete medium[51]. Unless specified, all chemicals and drugs were purchased from Sigma. Anti-HA (F-7) and anti-c-Myc (9E10) antibodies were obtained from Santa Cruz Biotechnology, whereas anti-Flag M2 antibody was purchased from Sigma. Glutathione–, Protein A– and Protein G–Sepharose beads were obtained from Amersham Pharmacia Biotech.

Strains and plasmids. *Escherichia coli* strains DH5α and DH10 were used for all plasmid maintenance and construction. Yeast strains used in this study are listed in Supplementary Table 1. Yeast transformation was performed using a standard LiCl method[52]. All *C. glabrata* deletion strains were derived from our wild-type laboratory strain BG2 or from a *ura3* derivative of BG2, BG14[53]. Gal4pDBD–Pdr1p or GST–Pdr1p fusion constructs were generated by cloning PCR fragments of *PDR1* of specified size and region, into pGBKT7 (Clontech) or pGEX2–TKN in frame with Gal4pDBD or GST, respectively, using *Nco*I and *Not*I restriction sites. Yeast constructs expressing C-terminal Flag-tagged full-length Gal11p (Gal11p–Flag$_2$) or Gal11p with deletion of the KIX domain (amino acids 1–100) (Gal11pΔKIX–Flag$_2$), and N-terminal Myc-tagged full-length wild-type Pdr1p (Myc$_6$–Pdr1pwt), were generated by subcloning the PCR products into the plasmids pCU416 and pCU315, respectively. DNA fragments encoding CgGal11p KIX domains (CgGal11Ap (amino acids 1–86) and CgGal11Bp (amino acids 1–82)) were amplified by PCR from genomic DNA and subcloned into pGEX–2TKN. Point-mutated pGEX–2TKN–Gal11pKIX (amino acids 1–100) plasmids were produced using the Quickchange XL site-directed mutagenesis kit (Stratagene). The mammalian constructs expressing HA–Gal4DBD activation domains of ScPdr1p, ScPdr3p, ScGal4p and CgPdr1p were generated by subcloning the PCR products into pcDNA3–HA–Gal4DBD using *Bam*HI and *Eco*RI sites. All plasmids were confirmed by DNA sequencing.

C. glabrata strain construction. The DNA sequences of the *C. glabrata* genes *PDR1* (CAGL0A00451g), *GAL11A* (CAGL0H06215g), *GAL11B* (CAGL0F00803g) and *MED1* (CAGL0D01386g) were taken from the Génolevures website (http://cbi.labri.fr/Genolevures/blast.php) and used to design primers for construction of disruption constructs and restoration constructs. Disruption constructs for *C. glabrata* genes are derived from pAP599 in which an *hph* expression cassette (from 5′ to 3′, a 500 base pair (bp) *S. cerevisiae PGK1* promoter followed by *Klebsiella pneumoniae hph* coding sequences (CDSs) and a 400 bp *S. cerevisiae HIS3* 3′ untranslated region (UTR)) that confers hygromycin B resistance (Hyg^R) is immediately flanked by *S. cerevisiae FRT* sites and then by multiple cloning sites (MCSs). A 0.5–1 kilobase (kb) 5′ UTR and a 0.5–1 kb 3′ UTR fragment of the target *C. glabrata* gene were amplified from BG2 genomic DNA by PCR and individually subcloned into pAP599. The accuracy of the cloned fragments was verified by DNA sequencing. The disruption construct containing the target gene 5′ and 3′ UTRs flanking the *hph* expression cassette was released from the gene disruption plasmid by restriction digest and used to transform BG14 to Hyg^R. The correct gene disruption was confirmed by PCR amplification. The *hph* expression cassette was then removed by transformation of the strains with a self-replicating plasmid, pRD16, that carries an expression cassette for *S. cerevisiae FLP1* (a 2.5 kb *C. glabrata EPA1* promoter followed by the *S. cerevisiae FLP1* coding region). The Flp1 recombinase recognizes the *FRT* sites immediately flanking the integrated *hph* expression cassette and releases the cassette. pRD16 was subsequently lost by streaking cells on plates containing 5-FOA, which selects against *URA3*. The resulting strains (BG1710–1713) were then restored to Ura$^+$ by transforming yeast cells with *Pst*I-digested pBC34.1, which carries an intact *C. glabrata URA3* on a 4.1 kb Pst1 fragment to generate strains 1718–1721. *gal11A gal11B* double mutant was generated by disruption of *GAL11B* in strain BG1710.

Treatment of yeast cultures with different xenobiotics. Yeast cultures were grown overnight with agitation in YPD at 30 °C. The next day, cells were pelleted and washed twice with sterilized MilliQ purified (Millipore) water. Cells were then resuspended in 1% yeast extract and 2% peptone (YP) to an optical density (OD$_{600}$) of 0.8, grown for another 16 h at 30 °C, then treated with different drugs for 20 min for quantitative RT–PCR, or 5 h for β-galactosidase assay.

Quantitative real-time RT–PCR. Total RNA was extracted from 5 ml of yeast culture using the Qiagen RNeasy MiniKit. Two micrograms of total RNA was used to generate cDNA using the First-Strand cDNA Synthesis Kit (GE Healthcare). The cDNA mix was diluted tenfold, and 2 μl was used for real-time quantitative PCR with SYBR Green (Applied Biosystems) on an ABI Prism 7900HT Sequence Detection System (Applied Biosystems).

β-Galactosidase assay. Yeast culture (1.5 ml) was harvested by brief spinning. The pellet was resuspended in 300 μl of Z buffer (10 mM Na$_2$HPO$_4$.7H$_2$O, 10 mM NaH$_2$PO$_4$.H$_2$O, 10 mM KCl, 1 mM MgSO$_4$.7H$_2$O, pH 7.0). One hundred microlitres of this suspension was transferred to a fresh tube and subjected to three cycles of freeze (30 s in liquid N$_2$) and thaw (1 min at 37 °C). After this, 0.7 ml of Z buffer with β-mercaptoethanol (0.27 ml β-mercaptoethanol in 100 ml of Z buffer) was added, immediately followed by addition of 160 μl of ONPG solution (4 mg ml^{-1} in Z buffer). The tube was kept at 30 °C until yellow colour developed, when 0.4 ml of 1 M Na$_2$CO$_3$ was added to stop the reaction. The cell debris was pelleted by centrifugation, and the supernatant was used to measure OD$_{420}$. All the readings were normalized to the concentration of protein (as determined by Bradford). All experiments were performed with three independent replicates.

Drug sensitivity assays. Fresh *S. cerevisiae* colonies were inoculated in 5 ml of YPD or SD selection medium, and grown overnight at 30 °C. The cells were diluted to an OD$_{600}$ of 0.2, and 2 μl was spotted on solid medium containing an increasing gradient of drugs (ketoconazole, cycloheximide or 4-NQO) in the agar. The plates were incubated at 30 °C for 2–3 days. Susceptibility of the *C. glabrata* mutants was tested by spotting serial dilutions of cells onto YPD agar plates supplemented with fluconazole or ketoconazole. The yeast strains were grown overnight at 30 °C in YPD liquid medium. Cells were diluted to an OD$_{600}$ of 2.0 in PBS, and 4 μl of the cell suspension and tenfold serial dilutions of the cells were spotted onto plates. Growth was assessed after 1 day of incubation at 30 °C.

Immunoprecipitation and immunoblotting. For co-immunoprecipitation, yeast cells were transformed with Myc$_6$–Pdr1pwt and Gal11p–Flag$_2$ (or Gal11pΔKIX–Flag$_2$). Cells were washed with 1 × PBS after overnight culture in YPD medium, and then cultured in YP for another 24 h. Cells were then treated with different concentrations of ketoconazole and cycloheximide (or ethanol for vehicle controls) for 1 h. Yeast cell lysates were extracted in immunoprecipitation buffer (50 mM Tris-HCl, pH 8.0, 140 mM NaCl, 0.1 mM EDTA, 10% glycerol, 0.02% NP-40, 1 mM DTT, 0.25 mM PMSF, 1 mM benzamidine, 0.5 mg ml^{-1} aprotinin and Protease Inhibitor Cocktail (Complete, Roche)), with or without xenobiotics in the presence of glass beads. The Flag-tagged proteins were then immunoprecipitated by adding anti-Flag M2 antibody-beads (Sigma) and incubating for 3 h with nutating at 4 °C. After washing five times with 1 ml of 0.25 M KCl IP buffer, bound proteins were eluted with 0.1 mg ml^{-1} Flag peptide (Sigma). Xenobiotics at the indicated concentrations were present at all steps of the co-immunoprecipitation, including washing. Immunoblotting was performed according to standard protocols. To immuno-purify Myc$_6$–Pdr1pwt or HA–Pdr1pwt, yeast whole-cell extract (in immunoprecipitation buffer) was incubated with anti-Myc (9E10) antibody or anti-HA (F-7) antibody, and 50% mix of Protein A– and G–Sepharose beads (Amersham) at 4 °C for 3 h. Beads were washed with IP buffer five times, and used for drug-binding experiments as outlined below.

Chromatin immunoprecipitation. Chromatin immunoprecipitation was performed according to standard procedures[54]. Briefly, yeast cells were grown to an OD$_{600}$ of 0.8 before fixing with 1% formaldehyde for 20 min. The cells were washed with Tris-buffered saline (20 mM Tris-HCl, pH 7.5, and 150 mM NaCl), and resuspended in 1 ml of FA lysis buffer (50 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate and 0.1% SDS). Cells were lysed using soda lime 0.5 mm glass beads and Beadbeater (BioSpec Products). The glass beads were then removed and the cell debris with chromatin was subjected to sonication (550 sonic dismembrator, Fisher) two times for 30 s each, separated by incubation for 2 min on ice. The soluble chromatin was collected in a fresh tube. Chromatin (300 μl) was used for immunoprecipitation with anti-HA Ab and Protein A–Sepharose beads (Amersham), and the immunoprecipitated chromatin was eluted with 200 μl of 0.5 mg ml^{-1} HA peptide (Sigma) in Tris-buffered saline. Input and immunoprecipitated chromatin were de-crosslinked at 65 °C overnight, and then passed through a Qiagen PCR purification column. The purified DNA was analysed by quantitative PCR in real-time using the ABI Prism 7900HT Sequence Detection System (Applied Biosystems). Relative occupancy values were calculated by determining the apparent immunoprecipitation efficiency and normalized to the level observed at an open reading frame (ORF)-free region from chromosome V.

Drug-binding and cold competition assay. Xenobiotic-binding experiments were performed essentially as described[55] with some modifications as explained below. Radiolabelled ketoconazole (KET: [^{3}H]G), specific radioactivity 10 Ci mmol^{-1}, was purchased from American Radiolabelled Chemicals. Beads with either immunopurified proteins or GST recombinant proteins were incubated with KET ([^{3}H]G) in drug-binding buffer (10 mM K$_2$HPO$_4$, 10 mM KH$_2$PO$_4$, pH 7.0, 2 mM EDTA, 50 mM NaCl, 1 mM DTT, 0.5 mM CHAPS, 10% glycerol and protease inhibitors) at 4 °C for 3 h with mixing in a volume of 500 μl. For cold competition, unlabelled xenobiotics were added to a concentration of 1,000-fold excess of KET. The beads were then washed thrice briefly in 900 μl of ice cold drug-binding buffer, carefully removing as much of each wash as

possible. The washed beads were resuspended in 100 μl of drug-binding buffer, added to 5 ml of scintillation fluid Aquasol-2 (Perkin Elmer) and mixed briefly by shaking. After 10 min, scintillation counting was performed in an LS6500 Multi-Purpose Scintillation Counter (Beckman Coulter). All experiments were performed with at least three replicates.

GST pull-down assays. Recombinant GST–fusion proteins were expressed in *E. coli* (BL21, DE3) and purified using glutathione–Sepharose beads according to standard protocol (Pharmacia). Beads with GST proteins were incubated with either whole-cell lysate or *in vitro* translated protein mix in binding buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1 mM EDTA, 10% glycerol, 0.05% NP-40, 1 mM DTT, 0.25 mM PMSF, 1 mM benzamidine and Protease Inhibitor Cocktail for yeast extracts (Sigma)), for 3 h at 4 °C. The beads were washed five times with wash buffer (20 mM Tris-HCl, pH 8.0, 250 mM KCl, 0.1 mM EDTA, 10% Glycerol, 0.1% NP-40, 1 mM DTT, 0.25 mM PMSF, 1 mM benzamidine and Protease Inhibitor Cocktail). The beads were finally washed once with bind-ing buffer. The bound proteins were then eluted with 50 μl of 0.3% sarkosyl in binding buffer for 1 h at 4 °C. The eluted proteins were resolved on 10% poly-acrylamide gel and detected either by immunoblotting or by autoradiography.

Large-scale purification and identification of GST–Gal11p KIX-associated proteins in yeast. Yeast (6 l) cultured in YPD medium overnight was harvested by centrifugation and washing once with distilled water. The cell pellet was then resuspended in 0.25 volumes of lysis buffer (50 mM Tris-HCl (pH 8.0), 400 mM NaCl, 5 mM MgCl₂, 1 mM EGTA, 1 mM EDTA, 0.1% NP40, 1 mM DTT, 0.25 mM PMSF, 1 mM benzamidine, 0.5 mg ml⁻¹ aprotinin and Protease Inhibitor Cocktail (Complete, Roche)), and the suspension was quick-frozen in liquid nitrogen. Frozen cells were lysed by grinding with a mortar and pestle together with dry ice. One volume of lysis buffer was added after evaporation with dry ice, and the extract was spun at 4,000*g* for 10 min. The supernatant was pre-incubated with 200 μl glutathione–Sepharose-bound GST for 2 h at 4 °C with rotation, then it was incubated for another 3 h at 4 °C with 200 μl glutathione–Sepharose-bound GST–Gal11p KIX (amino acids 1–100). The beads were washed seven times with wash buffer (20 mM Tris-HCl, pH 8.0, 250 mM KCl, 0.1 mM EDTA, 10% glycerol, 0.1% NP-40, 5 mM MgCl₂, 1 mM EGTA, 1 mM EDTA, 1 mM DTT, 0.25 mM PMSF, 1 mM benzamidine and Protease Inhibitor Cocktail). The beads were finally washed once with low salt (150 mM NaCl) wash buffer. The bound proteins were then eluted with 500 μL of 0.3% sarkosyl in binding buffer for 1 h at 4 °C and dialysed overnight in 1 l of dialysis buffer (1% SDS, 1 mM β-mercaptoethanol and 1 mM Tris-HCl, pH 8.0). The dialysed eluate was concentrated by dry-ice/ethanol SpeedVac to approxi-mately 80 μl. The eluted proteins were resolved on 10% polyacrylamide gel and stained with Coomassie colloidal blue. The specific band at about 120 kDa was excised and subjected to trypsin digestion, followed by liquid chromatography MS/MS (LC-MS/MS) at the Taplin Biological Mass Spectrometry Facility at Harvard Medical School.

***C. elegans* liquid killing assays.** The *C. glabrata* strains and lawns were grown and prepared as described previously[56]. The liquid medium killing assays were conducted as detailed in ref. 56, with a few changes. After the *glp-4;sek-1* worms were incubated at 25 °C for two days, they were washed off the nematode growth medium plates with M9 and transferred to *C. glabrata* lawns formed on Brain–Heart Infusion broth (Difco) agar plates. The worms were incubated on the lawns for 4 h at 25 °C, then washed off the plates with M9 buffer and allowed to crawl on unseeded Brain–Heart Infusion broth plates to remove yeast cells from their cuticles. Approximately 70–80 worms were then picked to wells in a six-well microtitre dish that contained 2 ml of liquid medium of 79% M9 buffer, 20% Brain–Heart Infusion broth, 10 μg ml⁻¹ cholesterol in ethanol, and 90 μg ml⁻¹ kanamycin, with the addition of 0.25% DMSO or 8 μg ml⁻¹ fluco-nazole. The plates were incubated at 25 °C overnight and then examined at 24 h intervals for survival. Worms were considered dead and removed when they did not respond to being touched by a platinum wire pick. *P* values were calculated based on the entire six-day experiment with the log-rank and Wilcoxon tests performed by STATA 6 statistical software (Stata).

Structural analysis by NMR. A pET24b plasmid containing the Gal11p KIX sequence with an N-terminal His₆-tag followed by a cleavage site for the tobacco etch virus protease was transformed into *E. coli* BL21 (DE3) cells. The cells were grown to an OD₆₀₀ of 0.7 at 37 °C and induced for 12–16 h at 25 °C with 1 mM isopropyl β-D-1-thiogalactopyranoside. The cells were lysed by sonication, cen-trifuged and the supernatant was purified by Ni-NTA resin (Qiagen). The His tag was cleaved off by Tev protease overnight at 4 °C. This sample was further purified by fast protein liquid chromatography using a size exclusion column (Sephadex 75, Pharmacia). All NMR samples were in PBS buffer (10 mM Na₂HPO₄, 2 mM K₂HPO₄, 137 mM NaCl, 2.7 mM KC1, 1 mM EDTA and

0.01% NaN₃), pH 6.5, unless otherwise stated. Pdr1pAD-34 was expressed as a GST-fusion tag in *E. coli* BL21 (DE3) cells. The protein was purified as described above with the fusion tag cleaved after Nickel resin purification. ¹⁵N/¹³C-labelled samples of Gal11p KIX and Pdr1pAD-34 were obtained by overexpression of the respective proteins in M9 minimal media enriched with ¹⁵NH₄Cl and/or [¹³C]glucose. Perdeuterated samples of Gal11p KIX were generated in similar fashion with the protein expressed in M9 minimal media in ²H₂O using [¹³C–²H]glucose. Pdr1pAD-12 (H₄N–EDLYSILWSDVY–COOH) and SREBP-1a-26 (H4N–EPCDLDAALLTDIEDMLQLINNQDSD–COOH) were purchased as synthetic peptides from Tufts New England Medical Center peptide synthesis facility. Titrations with the peptide were performed in a 'high phosphate buffer' (50 mM Na₂HPO₄, 50 mM NaH₂PO₄, 2 mM K₂HPO₄, 2.7 mM KCl ,1 mM EDTA and 0.01% NaN₃) to maintain the pH during the course of the titration. For NMR spectroscopy, backbone assignments were mostly obtained by the standard set of triple resonance experiments (HNCA/HNCOCA, HNCACB/CBCACONH, HNCO/HNCACO)[57]. Because the Gal11p KIX exhibited stretches that were difficult to assign, we used a time-shared (HA)CANH/(HN)NCAHA to verify and complete the backbone assignment[58]. The side-chain residues were assigned using HCCONH and CCONH experiments of Gal11p KIX in H₂O and HCCH-TOSY in ²H₂O. ¹⁵N-dispersed heteronuclear single-quantum coherence–nuclear overhauser enhancement spectroscopy (HSQC–NOESY) with a mixing time of 90 ms was recorded to provide distance constraints. All backbone experiments were performed on a Bruker 600 MHz spectrometer equipped with a cryoprobe. The side-chain and NOESY experiments were performed on Bruker 500 MHz and 900MHz spectrometers, respectively. Cross-saturation experiments were performed using perdeuterated samples of ¹⁵N-enriched Gal11p KIX and unlabelled Pdr1pAD-12. The methyl region (1 p.p.m.) was excited with a wideband uniform rate smooth truncation pulse for 2 s, and this was followed by a standard HSQC experiment. The experiment was performed in an interleaved fashion with irradiation and no irradiation in successive scans. The ratio of intensities between the irradiated HSQC and the non-irradiated HSQC was later analysed[59]. ARC105 KIX was expressed as a GST fusion protein and purified over a glutathione–Sepharose column. ARC105 KIX was cleaved from GST using tobacco etch virus protease. The cleaved protein was further purified by fast protein liquid chromatography using a size exclusion column (Sephadex 75, Pharmacia) in a buffer containing 30 mM NaCl, 5 mM Na₂HPO₄ and 5 mM KH₂PO₄ at pH 6.8. For titration experiments, unlabelled Pdr1pAD-34 was purified in the same buffer as mentioned above and added to ARC105 KIX.

For NMR Structure calculations, NMR structure refinement was performed using the program CYANA[60]. Torsion angle constrains (146 constrains) were calculated using the program TALOS[61]. Assignment of the backbone and side-chain resonances was performed using the program CARA. One thousand three hundred and fifty-nine distance restrains were calculated from the ¹⁵N-dispersed NOESY experiments. NOESY cross peaks were integrated using the program PeakInt.

51. Sherman, F., Fink, G. R. & Hicks, J. B. *Methods in Yeast Genetics* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1986).
52. Gietz, D., St Jean, A., Woods, R. A. & Schiestl, R. H. Improved method for high efficiency transformation of intact yeast cells. *Nucleic Acids Res.* **20**, 1425 (1992).
53. Cormack, B. P. & Falkow, S. Efficient homologous and illegitimate recombination in the opportunistic yeast pathogen *Candida glabrata*. *Genetics* **151**, 979–987 (1999).
54. Aparicio, J. G., Viggiani, C. J., Gibson, D. G. & Aparicio, O. M. The Rpd3-Sin3 histone deacetylase regulates replication timing and enables intra-S origin control in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **24**, 4769–4780 (2004).
55. Kepinski, S. & Leyser, O. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. *Nature* **435**, 446–451 (2005).
56. Breger, J. *et al.* Antifungal chemical compounds identified using a *C. elegans* pathogenicity assay. *PLoS Pathogens* **3**, e18 (2007).
57. Ferentz, A. E. & Wagner, G. NMR spectroscopy: a multifaceted approach to macromolecular structure. *Q. Rev. Biophys.* **33**, 29–65 (2000).
58. Frueh, D. P., Arthanari, H. & Wagner, G. Unambiguous assignment of NMR protein backbone signals with a time-shared triple-resonance experiment. *J. Biomol. NMR* **33**, 187–196 (2005).
59. Takahashi, H. *et al.* A novel NMR method for determining the interfaces of large protein–protein complexes. *Nature Struct. Mol. Biol.* **7**, 220–223 (2000).
60. Guntert, P., Mumenthaler, C. & Wuthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298 (1997).
61. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289–302 (1999).

LETTERS

A laser frequency comb that enables radial velocity measurements with a precision of 1 cm s⁻¹

Chih-Hao Li[1,2], Andrew J. Benedick[3], Peter Fendel[3,4], Alexander G. Glenday[1,2], Franz X. Kärtner[3], David F. Phillips[1], Dimitar Sasselov[1], Andrew Szentgyorgyi[1] & Ronald L. Walsworth[1,2]

Searches for extrasolar planets using the periodic Doppler shift of stellar spectral lines have recently achieved a precision of 60 cm s⁻¹ (ref. 1), which is sufficient to find a 5-Earth-mass planet in a Mercury-like orbit around a Sun-like star. To find a 1-Earth-mass planet in an Earth-like orbit, a precision of ∼5 cm s⁻¹ is necessary. The combination of a laser frequency comb with a Fabry–Pérot filtering cavity has been suggested as a promising approach to achieve such Doppler shift resolution via improved spectrograph wavelength calibration[2–4], with recent encouraging results[5]. Here we report the fabrication of such a filtered laser comb with up to 40-GHz (∼1-Å) line spacing, generated from a 1-GHz repetition-rate source, without compromising long-term stability, reproducibility or spectral resolution. This wide-line-spacing comb, or 'astro-comb', is well matched to the resolving power of high-resolution astrophysical spectrographs. The astro-comb should allow a precision as high as 1 cm s⁻¹ in astronomical radial velocity measurements.

The accuracy and long-term stability of state-of-the-art astrophysical spectrographs are currently limited by the wavelength-calibration source[6,7], typically either thorium–argon lamps or iodine absorption cells[8]. In addition, existing calibration sources are limited in the red-to-near-IR spectral bands most useful for exoplanet searches around M stars[9] and dark matter studies in globular clusters[10]. Iodine cells have very few spectral lines in the red and near-IR spectral bands, while thorium–argon lamps have limited lines and unstable bright features that saturate spectrograph detectors. Recently, laser frequency combs[11] have been suggested as potentially superior wavelength calibrators[2,3] because of their good long-term stability and reproducibility, and because they have useful lines in the red-to-near-IR range. The absolute optical frequencies of the comb lines are determined by $f = f_{ceo} + m \times f_{rep}$, where f_{rep} is the repetition rate, f_{ceo} is the carrier-envelope offset frequency and m is an integer. Both f_{rep} and f_{ceo} can be synchronized with radio-frequency oscillators referenced to atomic clocks. For example, using the generally available Global Positioning System (GPS), the frequencies of comb lines have long-term fractional stability and accuracy of better than 10^{-12}. For the calibration of an astrophysical spectrograph, fractional stability and accuracy of 3×10^{-11} are sufficient to measure a velocity variation of 1 cm s⁻¹ in astronomical objects. In addition, using GPS as the absolute reference allows the comparison of measurements at different observatories.

For existing laser combs, f_{rep} is usually <1 GHz (ref. 12), which would require a spectrograph with a resolving power of $R = \lambda/\Delta\lambda \gg 10^5$ to resolve individual comb lines (here $\Delta\lambda$ is the smallest difference in wavelengths that can be resolved at wavelength λ). In practice, astrophysical spectrographs tend to have a resolving power of
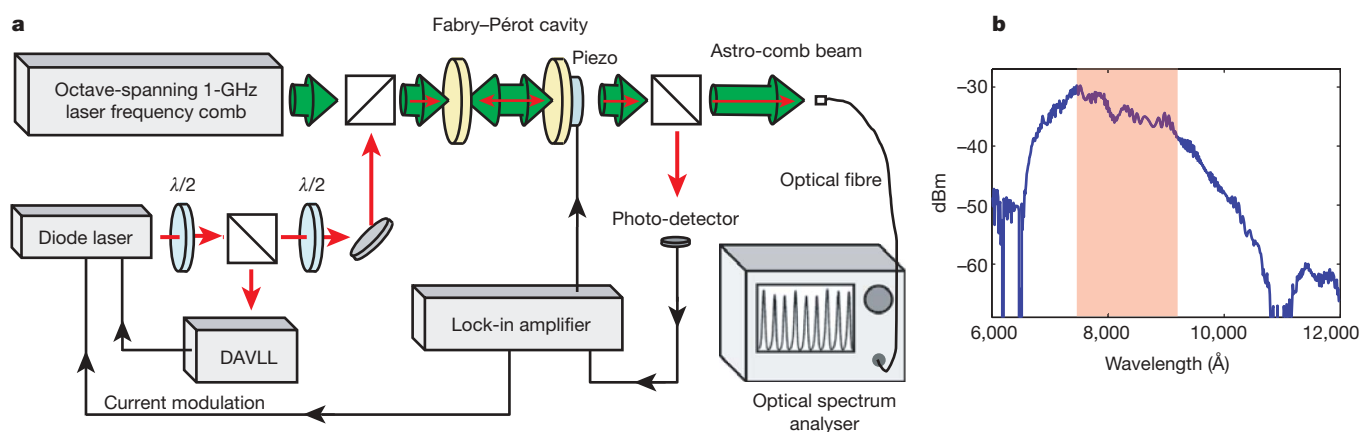
Figure 1 | Block diagram of the astro-comb. a, A stabilized 1-GHz frequency comb ('source-comb') using a mode-locked femtosecond laser passes through an FP cavity that filters out unwanted comb lines and increases the line spacing to at most 40 GHz (∼1 Å). For the demonstration spectra shown in this paper (**b**, and Fig. 2), the output beam from the astro-comb is collected by a single-mode fibre and measured using an optical spectrum analyser (Ando 6317) with resolution ∼8 GHz and reproducibility ∼2 GHz. **b,** Output spectrum of the 1-GHz source-comb. Typical operating parameters of the source-comb[13] are 600 mW of output power and an output spectrum from 6,000 Å to 12,000 Å with 9.3 W of pump power. The shaded area is the spectral range in which the current FP cavity mirrors have small GDD and hence provide good suppression of extraneous comb lines. The quantity dBm is ten times the logarithm of the power referenced to 1 mW.

[1]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, [2]Department of Physics, Harvard University, 17 Oxford Street, Cambridge, Massachusetts 02138, USA. [3]Department of Electrical Engineering and Computer Science and Research Laboratory for Electronics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [4]MenloSystems Inc., 69 Stickles Pond Road, Newton, New Jersey 07860, USA.

$R \approx 10^4$–10^5 owing to physical limitations on the instruments, including the telescope aperture, the grating collimator diameter and the grating blaze. Thus, a laser comb must have line spacing >10 GHz to serve as a practical wavelength calibrator. Therefore, we augmented a 1-GHz-repetition-rate laser comb with a stable broadband Fabry–Pérot (FP) cavity to increase the comb line spacing to 40 GHz over a range >1,000 Å. This novel[2], wide-line-spacing 'astro-comb' can provide improved wavelength calibration for a wide range of existing and planned astrophysical spectrographs.

The astro-comb set-up is shown schematically in Fig. 1. An octave-spanning optical frequency comb with a 1-GHz repetition rate ('source-comb') is generated by a mode-locked Ti:sapphire femto-second laser[13]. The linewidth of each comb line is <1 kHz, with both f_{rep} and f_{ceo} stabilized using low-noise frequency synthesizers, which can be referenced to an atomic clock. The stabilized source-comb light passes through an FP cavity that filters out unwanted comb lines and increases the line spacing. The FP cavity is stabilized by an injected diode laser signal that is itself stabilized to the Rb D1 line (7,947 Å) using a dichroic-atomic-vapour laser lock (DAVLL[14]).

To realize an astrophysical wavelength calibrator, the FP cavity must filter comb lines over a broad spectral range. The mirrors used in the plane-parallel FP cavity have ~99% reflectivity and optimized group delay dispersion (GDD) (<10 fs^2) in the range of 7,700 Å to 9,200 Å. We measured the finesse of the FP cavity to be ~250 at 7,947 Å, which is consistent with the theoretical limit estimated from the mirror reflectivity and Fresnel losses. The GDD-optimized mirrors enable the generation of a filtered comb spanning a bandwidth of ~1,000 Å. With straightforward adjustment of the free spectral range of the FP cavity to approximately equal an integer multiple of f_{rep}, we realized such comb-line filtering. For example, Fig. 2 shows the measured astro-comb output spectrum spanning a bandwidth of ~1,000 Å, with 37-GHz line spacing and power ~10–100 nW in each comb line. If the ratio of free spectral range to f_{rep} is not an integer, the span of the filtered comb lines is narrower and groups of filtered comb lines appear repeatedly within the bandwidth of the mirrors (Fig. 2c). This 'Vernier-like' pattern can be shifted in wavelength and modified by varying the source-comb f_{ceo} and f_{rep} or the free spectral range of the FP cavity. The adjustability of the astro-comb-line pattern may assist the calibration of spectrographs over the bandwidth of the mirrors.

In addition to tunable line spacing up to 40 GHz, appropriate for use with astrophysical spectrographs, the astro-comb exhibits the stability and extraneous-line suppression necessary in an improved wavelength calibrator. By comparisons with a hydrogen maser, we determined the frequency fractional stability of the source-comb (characterized by f_{rep} and f_{ceo}) to be better than 10^{-12} on timescales of seconds to hours. Ideally, the FP cavity changes only the amplitude of the astro-comb's output lines, and not their frequency. Thus, the required stability of the FP cavity is much less stringent than that of the source-comb. However, several source-comb lines lie inside the resolution bandwidth of a typical astrophysical spectrograph. Although the FP cavity has finite suppression of neighbouring comb lines, instability in it leads to changes in the line shape of the astro-comb output spectrum as measured by an astrophysical spectrograph. In Fig. 3, we show a direct measurement of the suppression of extraneous lines of the astro-comb. The measured single-sided suppression of extraneous comb lines of more than 25 dB is consistent with the measured FP cavity finesse of 250. The FP cavity is stabilized by locking one transmission resonance maximum to a DAVLL-stabilized diode laser. The DAVLL-stabilized FP cavity is quite robust, remaining locked for periods of days. The absolute uncertainty in the DAVLL stabilized system is below 0.5 MHz, which is more than sufficient to maintain a sensitivity of 1 cm s^{-1}. (As noted above, the required FP cavity stability is much less stringent than the

Figure 2 | Example astro-comb output spectrum with 37-GHz line spacing. **a**, The astro-comb is tuned to span a bandwidth of ~1,000 Å. The resolution of the figure is not high enough to show individual lines. The observed amplitude variation is primarily due to the amplitude variation of the source-comb, with a minor contribution from 'mismatch' of the source-comb lines and the transmission resonances of the FP cavity caused by the residual GDD of the mirrors (~0–5 fs^2). Small line shifts due to residual mirror GDD can be determined experimentally to high precision (<1 cm s^{-1}). For the TiO$_2$/SiO$_2$ multi-layer mirrors used in the FP cavity, line stability ~1 cm s^{-1} is expected on timescales of several years. **b**, A small portion of the full output spectrum, showing individual filtered comb lines. The width of the lines is set by the optical spectrum analyser's resolution (~8 GHz). **c**, Intentional mismatch between source-comb spacing f_{rep} and FP cavity free spectral range causes groups of filtered comb lines to appear repeatedly within the bandwidth of the mirrors. The prominent line at 7,947 Å is the injected diode laser signal used to stabilize the FP cavity.

Figure 3 | Suppression of extraneous source-comb lines for the astro-comb. Here, the FP cavity is tuned such that the astro-comb has a 25-GHz line spacing. A fast photo-diode and spectrum analyser are used to measure the power in the 24-GHz and 25-GHz beatnotes from the 1-GHz source-comb (**a**) and the 25-GHz astro-comb (**b**) in the band 8,150–8,450 Å. The ~22-dB suppression of the 24-GHz beatnote in the astro-comb output signal corresponds to a ~25-dB single-sided suppression of extraneous comb lines, consistent with the FP cavity finesse of 250. When the astro-comb is used as a wavelength calibrator for an astrophysical spectrograph, the extraneous-line suppression will be regularly monitored.

required 10-kHz source-comb stability.) Residual frequency noise in the DAVLL is $<300\,\mathrm{kHz\,Hz}^{-1/2}$; corresponding to an amplitude fluctuation of $<0.1\%$ for the suppressed (extraneous) comb lines. The resultant frequency noise in the desired astro-comb line spacing is $<3\,\mathrm{kHz\,Hz}^{-1/2}$. Taking advantage of the ultrastable source-comb lines, the astro-comb output spectrum measured by the spectrograph is more stable than the FP cavity by more than two orders of magnitude. Consequently, the stability of the astro-comb is more than adequate for wavelength calibration of astrophysical spectrographs to $1\,\mathrm{cm\,s}^{-1}$ sensitivity.

In May 2008, we will deploy an astro-comb wavelength calibrator at the Multiple Mirror Telescope (MTT) on Mt Hopkins, Arizona. We will demonstrate the ability of the astro-comb to calibrate the Hectochelle multi-object echelle spectrograph[15] in a 150-Å bandwidth around 8,500 Å, which will be especially useful for the study of dark matter and other phenomena in globular clusters. Here we estimate the expected wavelength calibration precision of the astro-comb over a typical 10-hour MMT/Hectochelle measurement. The 1-GHz source-comb, referenced to GPS, will have spectral lines with accuracy and long-term fractional stability better than 10^{-12}. The FP cavity will use mirrors (Lambda Research) with 99% reflectivity and minimal GDD ($<1\,\mathrm{fs}^2$ over the 150-Å bandwidth of one Hectochelle order). The free spectral range of the FP cavity will be set to $\sim25\,\mathrm{GHz}$, which maximizes the calibration sensitivity[2]. Residual FP cavity mirror GDD and fluctuations of the FP cavity resonance[16] (width $\sim150\,\mathrm{MHz}$) will lead to changes in the extraneous-line suppression, which is typically $\sim4\times10^{-3}$ (that is, 25 dB), of up to 0.2%. Therefore, an upper-limit estimate of the uncertainty of astro-comb-line centres is $(4\times10^{-3})\times0.2\%\times1\,\mathrm{GHz}\approx8\,\mathrm{kHz}$. This uncertainty results in a systematic error in astrophysical velocity measurements of approximately $(8\,\mathrm{kHz}/377\,\mathrm{THz})\times(3\times10^{10}\,\mathrm{cm\,s}^{-1})<1\,\mathrm{cm\,s}^{-1}$. In practice, the precision of Doppler-shift/redshift measurements will also be affected by telescope instability and astronomical light-source fluctuations[6].

Beyond our first demonstration, astro-combs should enable many observations that have previously been considered technically unachievable. One example is the search for a 1-Earth-mass planet in an Earth-like orbit around a Sun-like star, which requires a sensitivity of $5\,\mathrm{cm\,s}^{-1}$ and stability on at least a 1-year timescale. In 2009 or 2010 we will deploy an astro-comb at the HARPS-NEF (High-Accuracy Radial-velocity Planet Searcher of the New Earths Facility) spectrograph ($R=120{,}000$) being built by the Harvard Origins of Life Initiative for the William Herschel telescope to search for exoplanets. We will use broadband mirrors with 99% reflectivity and optimal GDD in the FP cavity to generate stable calibration lines in appropriate spectral bands, which we expect will make the HARPS-NEF spectrograph sensitive enough to find Earth-like planets. Additional wavelength coverage can also be realized by frequency doubling the source-comb. Another possible application of the astro-comb is the Sandage–Loeb test[17,18], a direct measurement of the decelerating expansion of the early Universe. This test requires an observation period of $>10\,\mathrm{years}$ with existing wavelength calibrators, but should be feasible with an observation period of $\sim3\,\mathrm{years}$ using astro-combs. Thus, by enabling a velocity-shift

preci>sion of $\sim1\,\mathrm{cm\,s}^{-1}$, broad wavelength coverage and reproducibility over many years and between telescopes, astro-combs should revolutionize astrophysical spectroscopy.

1. Lovis, C. et al. An extrasolar planetary system with three Neptune-mass planets. Nature 441, 305–309 (2006).
2. Murphy, M. T. et al. High-precision wavelength calibration with laser frequency combs. Mon. Not. R. Astron. Soc. 380, 839–847 (2007).
3. Schmidt, P. O., Kimeswenger, S., Kaeufl, H. U. A new generation of spectrometer calibration techniques based on optical frequency combs. In Proc. 2007 ESO Instrument Calibration Workshop (ESO Astrophysics Symposia series, Springer, in the press); preprint at ⟨http://arxiv.org/abs/0705.0763v1⟩.
4. Araujo-Hauck, C. et al. Future wavelength calibration standards at ESO: the laser frequency comb. ESO Messenger 129, 24–26 (2007).
5. Osterman, S. et al. A proposed laser frequency comb based wavelength reference for high resolution spectroscopy. Proc. SPIE 6693, 66931G-1–9 (2007).
6. Lovis, C. et al. The exoplanet hunter HARPS: unequalled accuracy and perspectives toward 1 cm s−1 precision. Proc. SPIE 6269, 62690P-1–23 (2006).
7. Udry, S. et al. The HARPS search for southern extra-solar planets. XI. Super-Earths (5 and 8 M⊕) in a 3-planet system. Astron. Astrophys. 469, L43–L47 (2007).
8. Butler, R. P. et al. Attaining Doppler precision of 3 m s−1. Publ. Astron. Soc. Pacif. 108, 500–509 (1996).
9. Tarter, J. et al. A reappraisal of the habitability of planets around M dwarf stars. Astrobiology 7, 30–65 (2007).
10. Storm, J., Carney, B. W. & Latham, D. W. Distances and luminosities for RR Lyrae stars in M5 and M92 from a Baade-Wesselink analysis. Astron. Astrophys. 290, 443–457 (1994).
11. Udem, Th, Holzwarth, R. & Hänsch, T. W. Optical frequency metrology. Nature 416, 233–237 (2002).
12. Bartels, A., Gebs, R., Kirchner, M. S. & Diddams, S. A. Spectrally resolved optical frequency comb from a self-referenced 5 GHz femtosecond laser. Opt. Lett. 32, 2553–2555 (2007).
13. Benedick, A., Birge, J., Mücke, O. D., Sander, M. & Kärtner, F. X. Octave spanning 1 GHz Ti:sapphire oscillator for HeNe CH4-based frequency combs and clocks. In CLEO/Europe 2007 (Munich, 17–22 June, 2007) (IEEE, 2007); doi:10.1109/CLEOE-IQEC.2007.4386249.
14. Corwin, K. L., Lu, Z.-T., Hand, C. F., Epstein, R. J. & Wieman, C. E. Frequency-stabilized diode laser with the Zeeman shift in an atomic vapor. Appl. Opt. 37, 3295–3298 (1998).
15. Szentgyorgyi, A. et al. Hectochelle: a multi-object echelle spectrograph for the converted MMT. Proc. SPIE 3355, 242–252 (1998).
16. Reeves, J. M., Garcia, O. & Sackett, C. A. Temperature stability of a dichroic atomic vapor laser lock. Appl. Opt. 45, 372–376 (2006).
17. Sandage, A. The change of redshift and apparent luminosity of galaxies due to the deceleration of selected expanding universes. Astrophys. J. 136, 319–333 (1962).
18. Loeb, A. Direct measurement of cosmological parameters from the cosmic deceleration of extragalactic objects. Astrophys. J. 499, L111–L114 (1998).

Author contributions D.S., A.S. and R.L.W. initiated the project. C.-H.L., D.F.P. and R.L.W. designed the comb filtering strategy. C.-H.L. and D.F.P. designed the Fabry–Pérot stabilization strategy. R.L.W. supervised the project. C.-H.L. constructed the comb filtering system and acquired the filtered-comb data. A.J.B., P.F. and F.X.K. provided the 1-GHz laser comb and assisted C.-H.L. in its integration with the comb filtering system. C.-H.L. performed the initial data analysis. C.-H.L., A.G.G., F.X.K., D.F.P., D.S., A.S. and R.L.W. interpreted the data and assessed systematic effects. D.S. and A.S. advised on the astrophysical relevance of the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.L.W. (rwalsworth@cfa.harvard.edu).

LETTERS

Superinsulator and quantum synchronization

Valerii M. Vinokur[1], Tatyana I. Baturina[1,2,3], Mikhail V. Fistul[4], Aleksey Yu. Mironov[2,3], Mikhail R. Baklanov[5]
& Christoph Strunk[3]

Synchronized oscillators are ubiquitous in nature[1], and synchronization plays a key part in various classical and quantum phenomena. Several experiments[2–4] have shown that in thin superconducting films, disorder enforces the droplet-like electronic texture—superconducting islands immersed into a normal matrix—and that tuning disorder drives the system from superconducting to insulating behaviour. In the vicinity of the transition, a distinct state[4] forms: a Cooper-pair insulator, with thermally activated conductivity. It results from synchronization of the phase of the superconducting order parameter at the islands across the whole system[5]. Here we show that at a certain finite temperature, a Cooper-pair insulator undergoes a transition to a superinsulating state with infinite resistance. We present experimental evidence of this transition in titanium nitride films and show that the superinsulating state is dual to the superconducting state: it is destroyed by a sufficiently strong critical magnetic field, and breaks down at some critical voltage that is analogous to the critical current in superconductors.

We consider an exemplary tunable system for the superconductor-to-insulator transition studies, an array of small superconducting islands, each coupled to its nearest neighbours by Josephson weak links[6] (Fig. 1). The behaviour of the array is quantified by two competing energy scales: E_J, the Josephson coupling energy of the two adjacent superconducting islands, and the charging energy E_c, the energy cost to transfer a charge $2e$ between the neighbouring islands. Depending on the ratio E_J/E_c, the system can either be in the superconducting state, $E_J > E_c$, or in the insulating state, $E_J < E_c$. Near the superconductor-to-insulator transition the dynamic screening decreases the charging energy. From the quantum uncertainty principle, $\tau_0 \Delta \approx RC\Delta \approx \hbar$, where τ_0 is the island charge lifetime, R is the leakage resistance, C is the capacitance, and Δ is the superconducting gap. By introducing the dimensionless conductance $g = 2\pi\hbar/(e^2 R)$,

we obtain the Coulomb blockade energy $E_c \approx e^2/C \approx \Delta/g$. In an array of superconducting islands the superconductor-to-insulator transition occurs in a region where $g \gtrsim 1$ (ref. 7), so we conclude that the relevant charging energy $E_c < \Delta$, and thus the charge transfer is mediated by the activated motion of Cooper pairs[8]. Furthermore, because $g \gtrsim 1$, a propagating Cooper pair spreads over several islands to form a charge soliton, an ultimate charge carrier for the thermally activated conductivity.

In the superconducting state, the array undergoes the Berezinskii–Kosterlitz–Thouless (BKT) transition, separating a superconducting low-temperature phase with the bound vortex–antivortex pairs from a resistive high-temperature phase with free vortices, the transition temperature being $T_{BKT} \cong E_J/k_B$. In the insulating state the charge binding–unbinding transition that is dual to the superconducting BKT transition occurs at temperature $T \cong E_c/k_B$ (refs 9 and 10). Here we show that this transition separates a low-temperature superinsulating phase with practically infinite resistance, and an insulator with the large but finite thermally activated resistance $R \propto \exp[\Delta_c/(k_B T)]$, where the collective Coulomb barrier $\Delta_c = E_c(L/d)$ in one-dimensional arrays and $\Delta_c = E_c\ln(L/d)$ in two-dimensional (2D) arrays (L being the characteristic linear size of the system and d being the size of the elemental cell of the array). These formulae hold provided the screening length in the system that is related to capacitance to the ground exceeds the sample size, which we assume to be the case. Generalizing the technique developed in ref. 5 for temperature interval $E_c/k_B < T < \Delta_c/k_B$ to the low-temperature case ($T < E_c/k_B$; see Methods), we find the low bias ($eV < \Delta_c$) current–voltage characteristic in a superinsulating state to be:

$$I_s \cong I_c \exp\left\{ -\frac{(\Delta_c - eV)^2 \exp(E_c/2k_B T)}{E_c \Delta_c} \right\} \qquad (1)$$

with I_c being the Josephson critical current.

To gain an insight into the transition from activated insulator to superinsulator, we notice that the conductivity of the Cooper-pair insulator is proportional to the concentration, $\aleph_s$, of thermally activated charge solitons. We introduce the local charge density, $n_s(\mathbf{r})$, which is normalized to give the soliton energy as $\Delta_c = E_c \int d\mathbf{r}\, n_s^2(\mathbf{r})$. The probability for such a local density to appear at point $\mathbf{r}$ is proportional to $\exp[-n_s^2(\mathbf{r})/(2\langle\delta n^2\rangle)]$, where $\langle\delta n^2\rangle$ is the mean square fluctuation of the local charge density. Thus, the soliton density, which is proportional to the probability that a soliton will appear, is a product of all these local probabilities at all the points of the system: $\aleph_s \propto \Pi_{\mathbf{r}}\exp[-n_s^2(\mathbf{r})/(2\langle\delta n^2\rangle)] = \exp\{-[1/(2\langle\delta n^2\rangle)]\int d\mathbf{r}\, n_s^2(\mathbf{r})\} = \exp\{-[\Delta_c/(2E_c\langle\delta n^2\rangle)]\}$. At temperatures above the charge binding–unbinding superinsulating transition, $T_{SI} \cong E_c/k_B$, the solitons are unbound and the equipartition theorem $\langle\delta n^2\rangle = k_B T/E_c$ gives rise to the thermally activated resistance $R \propto \exp(\Delta_c/k_B T)$. At low temperatures $T < T_{SI}$, the charge solitons and antisolitons are bound, and
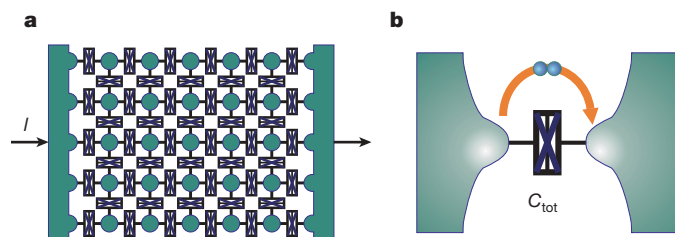
Figure 1 | Two-dimensional Josephson junction array. a, Sketch of the 2D array. Circles represent superconducting islands and crossed rectangles stand for Josephson weak links connecting these islands. The bias current I is injected to the left electrode and collected from the right electrode of the array. **b,** Phase synchronization allows us to view the 2D Josephson junction array as a single effective junction with the effective capacitance $C_{tot} = C/\ln N$.

[1]Material Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. [2]Institute of Semiconductor Physics, 630090, Novosibirsk, Russia. [3]Institut für experimentelle und angewandte Physik, Universität Regensburg, D-93040 Regensburg, Germany. [4]Theoretische Physik III, Ruhr-Universität Bochum, D-44801 Bochum, Germany. [5]IMEC, Kapeldreef 75, B-3001 Leuven, Belgium.

therefore $\langle \delta n^2 \rangle^{1/2}$ is the probability of breaking these pairs, that is, $\exp(-E_c/k_B T)$. This yields a double-exponential resistivity in the superinsulating phase:

$$R \propto \exp\left\{ \frac{\Delta_c}{E_c} \exp\left(\frac{E_c}{2k_B T} \right) \right\} \qquad (2)$$

This result follows from equation (1) on taking the limit $eV \ll \Delta_c$.

The physics of the charge-soliton-mediated transport stems from the quantum uncertainty principle $\Delta\varphi\Delta n \gtrsim 1$, where φ is the Josephson phase difference across the junction, and n is the number of the Cooper pairs transferred through the junction. Thus, charge and phase cannot be specified simultaneously to an arbitrary precision. The precise control of the charge at each junction enforced by the Coulomb blockade causes the corresponding phases φ at different junctions to fluctuate almost independently. However, when activated current passes through the system, the phases across the neighbouring junctions tend to synchronize each other to minimize Joule losses. The synchronization of phases of the superconducting order parameter across the system implies that the whole array can be viewed as a single effective junction with total capacitance C_{tot}; see Fig. 1b.

The Cooper pair is transferred across this effective junction in a form of a charge soliton spread over the array[9–12]. The one-dimensional array is a series of capacitances and $C_{tot}^{-1} = NC^{-1}$, where C is the capacitance between the adjacent islands and $N = L/d$ is the total number of islands. In the 2D case, taking into account that bias is applied from left to right, we find $C_{tot}^{-1} = 1/(2C) \ln N$, resulting in $\Delta_c = (E_c/2)\ln N$. This gives rise to the activation temperature dependence for the resistance, $R \propto \exp(\Delta_c/k_B T)$, in the temperature interval $E_c/k_B < T < \Delta_c/k_B$, where charge transport is mediated by the gas of free solitons and antisolitons (charge vortices). It yields the double-exponential resistivity of equation (2) at $T < E_c/k_B$, below the charge binding transition. The logarithmic scaling of the activation energy with the sample size agrees very well with the experimental findings[5].

The double-exponential current–charge (I–V) characteristic is derived for a regular array of Josephson junctions with all junction parameters identical. To examine the effect of irregularity in real systems we consider a one-dimensional array with position-dependent capacitances. Writing $C_{tot}^{-1} = \sum_i^N C_i^{-1} = N[(1/N)\sum_i^N C_i^{-1}] \equiv N\langle C^{-1} \rangle$, we obtain all the results for a regular array by substituting $E_c \rightarrow \langle E_c \rangle$. A similar consideration applies to 2D arrays. We thus conclude that the results obtained for regular arrays hold for systems with the random parameters, provided that the average $\langle C^{-1} \rangle$ is well defined.

The superinsulating state is experimentally observed in disordered titanium nitride (TiN) films which, to a degree of disorder, are in the vicinity of the superconductor-to-insulator transition[4]. Near the transition the conductance g is of the order of unity, according to experimental data at high magnetic fields in which superconductivity is suppressed and the film behaves in a metallic way[13]. We performed voltage-biased two-probe differential conductance measurements using standard low-frequency lock-in techniques with an alternating voltage of 10 μV. Magnetic fields were applied perpendicularly to the film surface.

Shown in Fig. 2 are plots of the differential conductance versus applied direct voltage and the corresponding I–V characteristics obtained by integration of the dI/dV versus V curves. The 70 mK data reveal 'normal' insulator behaviour: the differential conductivity is finite (Fig. 2a), and the I–V dependence is linear (Fig. 2b) up to a direct voltage of 10^{-3} V, reflecting ohmic conductivity behaviour in the activated insulator state. Lowering the temperature down to 20 mK drives the film into a superinsulating state: the differential conductivity and current remain immeasurably small at low bias voltages. At the depinning threshold voltage V_T, dI/dV abruptly jumps up over at least four orders of magnitude. The high- and low-temperature $\log I$–$\log V$ curves nearly coincide at high voltages and dramatically diverge at the low bias, which indicates that the superinsulating transition at $B = 0.9$ T occurs somewhere in between 20 mK and 70 mK.

As we have shown above, the superinsulator critical temperature is $T_{SI} = E_c/(2k_B) \cong \Delta/(gk_B)$. The magnetic field suppresses the superconducting gap Δ, so the critical temperature depends on the magnetic field. This defines a superinsulating critical field B_{SI}. Thus the superinsulating transition can be crossed either by varying the temperature or by tuning the magnetic field. The fan-like set of $\log I$ versus $\log V$ curves in Fig. 3b offers unambiguous evidence of the

Current (A)

Figure 3 | Magnetic-field-tuned transition to superinsulating state. a, The two-dimensional colour map of the current values in the B–V plane. The colour scale on the right-hand side represents current. The black domain in the map corresponds to the superinsulating state. The border between the black and coloured areas represents the field dependence of the threshold voltage on the magnetic field. **b**, Fan-like $\log I$–$\log V$ curves at $T = 20$ mK. The critical field B_{SI} is crossed on decreasing the magnetic field. Two diverging families of $\log I$–$\log V$ curves represent the 'normal' insulator at $B \geq 2.15$ T and the superinsulator at $B \leq 2.05$ T, and the conductivity shows linear I versus V dependence at low biases in the 'normal' insulator state, and sharp depinning at the voltage threshold in the superinsulator.

Figure 2 | Conductivity in normal-insulating and superinsulating states. a, Differential conductance dI/dV versus direct voltage in the insulating (conductive) state at 70 mK and in the superinsulating (zero-conductive) state at $T = 20$ mK measured for magnetic field $B = 0.9$ T. **b**, The corresponding I–V characteristics plotted in a log–log scale. In the insulating state the I–V curve is linear (ohmic) at small voltages. In the superinsulating state ($T = 20$ mK) the I–V curve shows depinning behaviour and an unmeasurably small current below the threshold voltage. The transition from superinsulator to 'normal' insulator occurs somewhere within the temperature interval 20–70 mK.

614

Figure 4 | Sketch of dual-phase diagrams for a superconductor and a superinsulator. a, Magnetic-field–temperature–current (B–T–I) superconductor phase diagram. **b**, A dual-phase diagram for a superinsulator is obtained from the superconductor phase diagram by interchanging the I and V axes. The threshold voltage V_T, the maximal voltage at which a superinsulator can retain a zero-conductivity state, corresponds to the critical current of a superconductor. The critical temperature for a superconductor T_{SC} maps onto the critical temperature of the 'normal' insulator–superinsulator transition T_{SI}.

superinsulating transition, placing B_{SI} at $T = 20$ mK between 2.05 and 2.15 T. Figure 3a displays current I versus B and V on a colour map. The border enclosing the infinitely resistant superinsulating domain (in black) visualizes the dependence of the threshold voltage upon the magnetic field.

Figure 4 summarizes our findings and presents a sketch of the dual superconducting (coordinates B, T, I), and superinsulating (coordinates B, T, V) phase diagrams. It shows a mirror-like symmetry between the superconducting and superinsulating phases: both collective states occupy the low-magnitude corners of their respective phase diagrams. In both cases the relevant variables are the magnetic field and temperature, the current for a superconductor and the voltage for a superinsulator. The temperature dependence of the critical field B_{SI} of the superinsulator shown in Fig. 4b follows the temperature behaviour of the upper critical field B_{c2} of the superconductor (Fig. 4a), given that $T_{SI} \propto \Delta/k_B \propto T_c$.

In conclusion, we note two things. First, the origin of the duality between a superinsulator and a superconductor lies in the conjugation of superconducting phase φ and condensate charge $Q = 2en$ connected by the uncertainty principle $\Delta\varphi\Delta n \geq 1$, where n is the number of Cooper pairs involved in the elemental charge transfer process. The collective phase characterizing a superconductor maps to the collective charge of a superinsulator. As a result, the duality between these two macroscopic quantum phenomena manifests itself via the mapping of all the characteristic parameters: $E_c \leftrightarrow E_J$, $I \leftrightarrow V$, and resistivity$\leftrightarrow$conductivity. Further, the duality manifests itself in the mirror symmetry of the phase diagram of both states: the upper critical field B_{c2} of a superconductor has its counterpart in the critical field B_{SI} for a superinsulator. The latter depends on temperature similarly to B_{c2}, while the temperature and field dependencies of the superconducting critical current are mirrored by those of the threshold voltage for depinning. This dual similarity extends even further. The Joule loss $P = IV$, which is exactly zero in the superconducting state, is also exactly zero in the superinsulator. Whereas the absence of Joule loss in a superconductor is the result of the nondissipative flow of the current and thus the lack of the voltage drop $V = 0$, the zero Joule loss in a superinsulator is ensured by the absence of the current at $V < V_T$, where V_T is the threshold voltage.

Second, our theoretical results were derived for a regular array of Josephson junctions. However, the experiments revealing a superinsulating state were carried out on homogeneously disordered films rather than on the artificially designed Josephson junction patterns. Our understanding of the origin of the superinsulating state in the

films relies on the formation of the network of superconducting droplets within the normal matrix. This network of superconducting droplets is precisely the array of superconducting weakly coupled islands considered above, provided this network maintains a relatively regular structure. Although the analytical theory of the droplet state is unknown, we conjecture that the droplet state is an inherent property of the critical region of the superconductor-to-insulator transition in the films and that a regular droplet array may emerge analogously to nucleation of the superconducting vortex lattice on the other side of the transition.

METHODS SUMMARY

Our analytical derivation of the I–V characteristic is based on the general linear-response theory, giving the Josephson current across a system confined between the two superconducting leads as $I_s = \langle \partial H/\partial \varphi \rangle$, where φ is the phase difference between the electrodes, H is the hamiltonian of the system of interest, and the angle brackets denote quantum mechanical and thermodynamic averaging. Considering a Josephson junction array in the insulating phase, where $E_c > E_J$, and making use of the perturbation theory with respect to small E_J/E_c, we express the d.c. Josephson current as a Fourier transform of the time-dependent Josephson phase correlation function $K(t)$ across the array, where the applied bias voltage V plays the role of the Fourier parameter. The I–V characteristic of the Josephson junction array is determined by the instanton phase configuration in which superconducting phases at all Josephson junctions evolve in a synchronized manner.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Pikovsky, A., Rosenblum, M. & Kurths, J. *Synchronization: A Universal Concept in Nonlinear Science* (Cambridge Univ. Press, Cambridge, 2001).
2. Kowal, D. & Ovadyahu, Z. Disorder induced granularity in an amorphous superconductor. *Solid State Commun.* **90**, 783–786 (1994).
3. Gantmakher, V. F., Golubkov, M. V., Lok, J. G. S. & Geim, A. K. Giant negative magnetoresistance of semi-insulating amorphous indium oxide films in strong magnetic field. *Zh. Eksp. Teor. Fiz.* 109, 1765–1778 (1996); *JETP* **82**, 951–958 (1996).
4. Baturina, T. I., Mironov, A., Yu., Vinokur, V. M., Baklanov, M. R. & Strunk, C. Localized superconductivity in the quantum-critical region of the disorder-driven superconductor-insulator transition in TiN thin films. *Phys. Rev. Lett.* **99**, 257003 (2007).
5. Fistul, M. V., Vinokur, V. M. & Baturina, T. I. Collective Cooper-pair transport in the insulating state of Josephson junction arrays. *Phys. Rev. Lett.* **100**, 086805 (2008).
6. Tinkham, M. *Introduction to Superconductivity* 2nd edn, Ch. 6 (McGraw-Hill, New York, 1996).
7. Beloborodov, I. S., Fominov, Ya. V., Lopatin, A. V. & Vinokur, V. M. Insulating state of granular superconductors in a strong-coupling regime. *Phys. Rev. B* **74**, 014502 (2006).
8. Lopatin, A. V. & Vinokur, V. M. Hopping transport in granular superconductors. *Phys. Rev. B* **75**, 092201 (2007).
9. Mooij, J. E. *et al.* Unbinding of charge-anticharge pairs in two-dimensional arrays of small tunnel junctions. *Phys. Rev. Lett.* **65**, 645–649 (1990).
10. Fazio, R. & Schön, G. Charge and vortex dynamics in arrays of tunnel junctions. *Phys. Rev. Lett.* **43**, 5307–5320 (1991).
11. Averin, D. V. & Likharev, K. K. in *Mesoscopic Phenomena in Solids* (eds Altshuler, B. L. *et al.*) 173–271 (Elsevier, Amsterdam, 1991).
12. Altland, A., Glazman, L. I., Kamenev, A. & Meyer, J. S. Inelastic electron transport in granular arrays. *Ann. Phys.* **321**, 2566–2603 (2006).
13. Baturina, T. I., Strunk, C., Baklanov, M. R. & Satta, A. Quantum metallicity on the high-field side of the superconductor-insulator transition. *Phys. Rev. Lett.* **98**, 127003 (2007).

METHODS

Consider $N \times M$ superconducting islands comprising a one- ($M = 1$) or two-dimensional Josephson junction array placed between the two superconducting (left and right, see Fig. 1) leads and closed by a small (as compared to the quantum resistance for the Cooper pairs $R_{CP} = h/4e^2 \approx 6.45\,k\Omega$) external resistance R_{ext}. Assigning the superconducting phase $\chi_{ij}(t)$ to the island in the ith column and the jth row, we write the hamiltonian of the array[5]:

$$H = H_0 + H_{int} + \frac{\hbar^2}{8E_c} \sum_{j=1}^{M} \left(\dot{\chi}_{1j}(t) + \dot{\chi}_{Nj}(t) \right)^2 - 2E_J \sum_{j=1}^{M} \cos\left[\frac{\chi_{1j}(t) + \chi_{Nj}(t)}{2} \right] \times$$
$$\times \cos\left[\frac{2eVt/\hbar + \psi(t) + \chi_{1j}(t) - \chi_{Nj}(t)}{2} \right] \tag{3}$$

Here

$$H_0 = \sum_{\langle ij, kl \rangle} \left[\frac{\hbar^2}{4E_c} \left(\dot{\chi}_{ij} - \dot{\chi}_{kl} \right)^2 - E_J \cos\left(\chi_{ij} - \chi_{kl} \right) \right] + \sum_{ij} \frac{\hbar^2}{4E_{c0}} \dot{\chi}_{ij}^2 \tag{4}$$

and the brackets $\langle ij, kl \rangle$ denote summation over the pairs of adjacent junctions, and the last term in equation (4) represents the self-charge energies of superconducting islands. The H_{int} term in equation (3) describes the coupling of phases on the leads to the thermal heat bath[14]. The charging energy E_{c0} is related to the each island's capacitance to the ground C_0. The phases of the left and right leads, $\chi_L(t)$ and $\chi_R(t)$, are fixed by the d.c. voltage V across the array:

$$\chi_R - \chi_L = 2eVt/\hbar + \psi(t) \tag{5}$$

where $\psi(t)$ describes thermal fluctuations in the leads. In the hamiltonian we have singled out the leftmost ($i = 1$) and rightmost ($i = N$) columns of islands directly coupled (adjacent) to the left and right leads respectively. The d.c. Josephson current through the Josephson junction array is given by the standard expression[11]:

$$I_s(V) = \left\langle \frac{\partial H}{\partial[\chi_L - \chi_R]} \right\rangle \tag{6}$$

and acquires the form:

$$I_s(V) = I_c \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau dt \sum_{j=1}^{M} \left\langle \cos\left[\frac{\chi_{1j}(t) + \chi_{Nj}(t)}{2} \right] \sin\left[\frac{2eVt/\hbar + \psi(t) + \chi_{1j}(t) - \chi_{Nj}(t)}{2} \right] \right\rangle \tag{7}$$

where the angle brackets $\langle \ldots \rangle$ stand for averaging over thermal fluctuations in the leads, $\psi(t)$, over quantum mechanical averaging over phases of internal junctions, $\chi_{ij}(t)$, and over the variable $\phi_j = (\chi_{1j} + \chi_{Nj})/2$. In the insulator regime that we address here, both E_c and $E_{c0} \gg E_J$, and we can calculate the quantum-mechanical average $\langle \cos\phi_j \rangle$ in the first-order perturbation theory with respect to small E_J/E_c. In most experimental situations $C \gg C_0$, and thus $E_c \ll E_{c0}$, so we can safely neglect the last term in equation (4) and obtain[15]:

$$\left\langle \cos\phi_j \right\rangle = \frac{E_J}{2E_c} \cos\left[\frac{2eVt/\hbar + \psi(t) + \chi_1(t) - \chi_N(t)}{2} \right] \sum_{n=-\infty}^{\infty} \frac{\exp(-E_c n^2/(4k_B T))}{n^2 - 1/4} \cong$$
$$\cong \frac{E_J}{2E_c} \cos\left[\frac{2eVt/\hbar + \psi(t) + \chi_1(t) - \chi_N(t)}{2} \right] \tag{8}$$

Using the time-dependent part of the hamiltonian as a perturbation we calculate the d.c. current in the framework of the linear response theory[16]:

$$I_s(V) = 2MI_c \frac{E_J}{\hbar} \frac{E_J^2}{E_c^2} \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau dt \int_0^t ds \left\langle \left[\hat{F}(t-s), \hat{F}(s) \right] \right\rangle_{H_0, \psi} \tag{9}$$

where

$$F(t) = \sin\left[\frac{2eV}{\hbar} t + \psi(t) + \chi_{1j}(t) - \chi_{Nj}(t) \right] \tag{10}$$

and $\left[\hat{F}(t-s), \hat{F}(s) \right]$ is the commutator of the corresponding Heisenberg operators. The notation $\langle \ldots \rangle_{H_0, \psi}$ means that the quantum mechanical averaging is carried out with respect to the fluctuational field ψ and with the unperturbed hamiltonian $H = H_0$. Assuming gaussian current noise (Nyquist noise) in the leads, we obtain the final expression for the d.c. current:

$$I_s(v) = 4MI_c \frac{E_J}{\hbar} \frac{E_J^2}{E_c^2} \Im m \int_0^\infty dt \exp[-t\delta/\hbar + i(2eVt/\hbar)] K(t) \tag{11}$$

where $\delta = 4e^2 R_{ext} k_B T$ (see refs 14 and 17) and the correlation function $K(t)$ of the

internal phases is defined as:

$$K(t) = \left\langle \exp i \left[\chi_{1j}(t) - \chi_{1j}(0) - \chi_{Nj}(t) + \chi_{Nj}(0) \right] \right\rangle_{H_0} \tag{12}$$

For the two-junction system (a single Cooper pair transistor), $\chi_{1j} \equiv \chi_{Nj}$, thus $K(t) \equiv 1$, and we obtain the results of refs 18 and 19.

In the zero approximation over E_J/E_c we neglect the Josephson coupling inside the array, and $K(t)$ can be found in a closed form as an analytical continuation of $K(\tau)$, where τ is imaginary time[5,20]:

$$K(\tau) = \int D[\chi_{ij}] \exp\left\{ i \left[\chi_{1j}(\tau) - \chi_{1j}(0) - \chi_{Nj}(\tau) + \chi_{Nj}(0) \right] \right\} \times$$
$$\times \exp\left(-\frac{\hbar}{4} \int_0^{\hbar/(k_B T)} d\tilde{\tau} \left[\sum_{\langle ij, kl \rangle} \frac{\left(\dot{\chi}_{ij}(\tilde{\tau}) - \dot{\chi}_{kl}(\tilde{\tau}) \right)^2}{E_c} + \sum_{ij} \frac{\left(\dot{\chi}_{ij}(\tilde{\tau}) \right)^2}{E_{c0}} \right] \right) \tag{13}$$

To calculate $K(t)$, we note that the phases $\chi_{ij}(\tau)$ are the periodic functions of τ with the period $\hbar/(k_B T)$ and thus can be presented as $\chi_{ij} + 2\pi M_{ij}(k_B T\tau/\hbar)$, where χ_{ij} is now defined for the interval $0 < \tau < \hbar/k_B T$ and M_{ij} are the so-called (integer) winding numbers[20]. At low temperatures and under the condition $E_c \ll E_{c0}$, so that the last term in the exponent can be dropped, the functional integral (13) is determined by the instanton configuration where all $\chi_{ij}(\tau)$ evolve collectively, reflecting phase synchronization across the sample. By Fourier transformation, the functional integration is reduced to calculation of gaussian integrals, and the correlation function can be conveniently presented as an interpolation:

$$K(t) = \exp\left(-\frac{\Delta_c E_c \xi^2 t}{\hbar^2} - i\frac{2\Delta_c t}{\hbar} \right) \tag{14}$$

Plugging formula (14) into the general expression (11) for the current, we arrive at the I–V characteristic. The interpolation function $\xi(T) = 1$ in the temperature interval $E_c < k_B T < \Delta_c$, where all the non-zero winding numbers can be neglected[5]. At ultralow temperatures $k_B T < E_c$, where the phase quantization becomes essential, the non-zero winding numbers should be taken into account. In the one-dimensional case we can write the correlation function $K(t)$:

$$K(t) = \exp\left(\frac{iNE_c t}{2} \right) \left[Z^{-1} \sum_n \exp\left(iE_c nt - n^2 E_c/2k_B T \right) \right]^N \tag{15}$$

where n^2 appear as the quantum numbers of the free rotator and the normalization factor:

$$Z = \sum_n \exp\left(-\frac{n^2 E_c}{2k_B T} \right) \tag{16}$$

At $k_B T < E_c$ the main contribution comes from the winding numbers $n = 0$, -1, $+1$, and we obtain the correlation function as equation (14) with $\xi(T) = \exp(-E_c/k_B T)$ and the collective Coulomb barrier $\Delta_c = E_c (L/d)$, where L is the linear size of the array and d is the size of a single junction. The 2D situation is more involved, and we need to adopt the technique developed in ref. 21 for the superconducting BKT transition. We can then derive the expression (14) for the correlation function with the same exponential form of ξ and $\Delta_c = E_c \ln(L/d)$.

14. Ingold, G.-L. & Nazarov, YuV in *Single Charge Tunneling* (eds Grabert, H. & Devoret, M. H.) Vol. 294 21–107 (NATO ASI Series B, Plenum, New York, 1991).
15. Landau, L. D. & Lifshitz, E. M. *Quantum Mechanics (Non-Relativistic Theory)* Ch. 6 142–146 (Elsevier Science, Oxford, UK/ Burlington, Massachusetts, 2003).
16. Ingold. G.-L.. in *Quantum Transport and Dissipation* (eds Dittrich, T. *et al.*) Ch. 4 213–248 (Wiley-VCH, Weinheim, 1998).
17. Koval, Y., Fistul, M. V. & Ustinov, A. V. Enhancement of Josephson phase diffusion by microwaves. *Phys. Rev. Lett.* **93**, 087004 (2004).
18. Matveev, K. A., Gisselfält, M., Glazman, L. I., Jonson, M. & Shekhter, R. I. Parity-induced suppression of the Coulomb blockade of Josephson tunneling. *Phys. Rev. Lett.* **70**, 2940–2943 (1993).
19. Lotkhov, S. V., Bogoslovsky, S. A., Zorin, A. B. & Niemeyer, J. Cooper pair cotunneling in single charge transistors with dissipative electromagnetic environment. *Phys. Rev. Lett.* **91**, 197002 (2003).
20. Efetov, K. B. & Tschersich, A. Coulomb effects in granular materials at not very low temperatures. *Phys. Rev. B* **67**, 174205 (2003).
21. José, J. V., Kadanoff, L. P., Kirkpatrick, S. & Nelson, D. R. Renormalization, vortices, and symmetry-breaking perturbations in the two-dimensional planar model. *Phys. Rev. B* **16**, 1217–1241 (1997).

LETTERS

Dust–climate couplings over the past 800,000 years from the EPICA Dome C ice core

F. Lambert[1,2], B. Delmonte[3], J. R. Petit[4], M. Bigler[1,5], P. R. Kaufmann[1,2], M. A. Hutterli[6], T. F. Stocker[1,2], U. Ruth[7], J. P. Steffensen[5] & V. Maggi[3]

Dust can affect the radiative balance of the atmosphere by absorbing or reflecting incoming solar radiation[1]; it can also be a source of micronutrients, such as iron, to the ocean[2]. It has been suggested that production, transport and deposition of dust is influenced by climatic changes on glacial–interglacial timescales[3–6]. Here we present a high-resolution record of aeolian dust from the EPICA Dome C ice core in East Antarctica, which provides an undisturbed climate sequence over the past eight climatic cycles[7,8]. We find that there is a significant correlation between dust flux and temperature records during glacial periods that is absent during interglacial periods. Our data suggest that dust flux is increasingly correlated with Antarctic temperature as the climate becomes colder. We interpret this as progressive coupling of the climates of Antarctic and lower latitudes. Limited changes in glacial–interglacial atmospheric transport time[4,9,10] suggest that the sources and lifetime of dust are the main factors controlling the high glacial dust input. We propose that the observed ~25-fold increase in glacial dust flux over all eight glacial periods can be attributed to a strengthening of South American dust sources, together with a longer lifetime for atmospheric dust particles in the upper troposphere resulting from a reduced hydrological cycle during the ice ages.

The EPICA (European Project for Ice Coring in Antarctica) ice core drilled at Dome C (hereafter EDC) in East Antarctica (75° 06′ S; 123° 21′ E) covers the past 800,000 yr (Fig. 1a). The dust flux record of Vostok (Fig. 1b) is thus extended over four additional cycles (Fig. 1c). The glacial–interglacial climate changes are well reflected in the sequence of high and low dust concentrations with typical values from 800 to 15 µg kg^{-1} and a ratio of 50 to 1 over most of the past eight climate cycles. The concentration of insoluble dust in snow depends on a number of factors such as the primary supply of small mineral particles from the continents, which is related to climate and environmental conditions in the source region[11], the snow accumulation rate, the long-range transport, and the cleansing of the atmosphere associated with the hydrological cycle. The strontium and neodymium isotopic signature of dust[12] revealed that southern South America was the dominant dust source for East Antarctica during glacial times[13], although contributions from other sources are possible during interglacials[14]. Because of the low accumulation rate at Dome C (~3 cm yr^{-1} water equivalent), dry deposition is dominant and the atmospheric dust load is best represented by the dust flux[15]. The total dust flux and the magnitude of the glacial–interglacial changes are remarkably uniform within the East Antarctic Plateau, as shown by the similarity between the EDC and the Vostok records (Fig. 1b, despite some chronological differences

and a 10-times finer resolution at EDC) over the past four climatic cycles and also depicted by the Dome Fuji dust record[16] (not shown).

At EDC, interglacials display dust fluxes similar to that of the Holocene (~400 µg m^{-2} yr^{-1}). However, some differences can be seen in the record before and after the Mid-Brunhes Event (MBE, ~430 kyr BP) which is considered a transition in the climatic record[7,8] from cooler (for example Marine Isotopic Stages, MIS 13, 15, 17) to warmer (for example MIS 11, 9 and 5.5) interglacials. Before the MBE there were fewer occurrences of low concentrations, and warm periods represent ~12% of the time, compared with ~30% after the MBE. All eight glacial periods appear similar in magnitude and show an average increase in dust flux by a factor of about 25, with glacial maxima displaying fluxes of at least 12 mg m^{-2} yr^{-1}. The weakest glacial stages in the EDC ice core are MIS 14 and 16. For MIS 14 this is consistent with the findings from terrestrial and marine records[6,17,18]. In contrast, MIS 16 in those records is the strongest glacial in the Late Quaternary period.

The extension of the EDC dust record to 800 kyr BP confirms the increased atmospheric dust load during cold periods of the Quaternary period with respect to warm stages. The first-order similarity of EDC dust with the global ice-volume record (Fig. 1e, $r^2 = 0.6$) confirms that major aeolian deflation in the Southern Hemisphere was linked to Pleistocene glaciations. Comparison with the magnetic susceptibility record of loess/palaeosol sequences from the Chinese Loess Plateau (Fig. 1f) also provides evidence for broad synchronicity of global changes in atmospheric dust load.

EDC dust has been measured using both a Coulter counter and a laser sensor (see Methods). Laser measurements are obtained at higher resolution along the core, but dust size is difficult to calibrate; therefore only the relative variation of the signal is used. Overall, the Coulter counter and laser (relative) size records (Fig. 1d) are in good agreement. The slight discrepancies during MIS 5.5, 6 and 12 are possibly related to the different sampling resolution, as size data from the laser and Coulter counter represent a continuous 1.1-m average and discrete 7-cm subsamples every 0.5–6 m, respectively. From MIS 14 (~2,900 m depth) and downwards in the ice core, the dust size profile is not available because of the presence of particle aggregates formed in the ice. This phenomenon, which needs further investigation, has been observed in the EDC core only in very deep glacial sections, where ice thinning becomes very important and *in situ* temperature higher than −8 °C may allow partial melting around particles. This problem was solved through sonication of the samples, which allowed us to obtain reliable concentration data (see Methods). For the upper part of the record, larger (smaller) particles are generally observed during warm (cold) periods, as reflected by the

Figure 1 | EDC dust data in comparison with other climatic indicators.
a, Stable isotope (δD) record from the EPICA Dome C (EDC) ice core[8] back to Marine Isotopic Stage 20 (EDC3 timescale) showing Quaternary temperature variations in Antarctica. **b**, Vostok dust flux record (Coulter counter) plotted on its original timescale[11]. **c**, EDC dust flux records. Red and grey lines represent, respectively, Coulter counter (55-cm to 6-m resolution) and laser-scattering data (55-cm mean). Numbers indicate Marine Isotopic Stages. Note that the vertical extent of the scales of **b** and **c** is larger than for the other records. **d**, EDC dust size data expressed as FPP (see Methods). The orange and grey curves represent measurements by Coulter counter (2-kyr mean) and laser (1-kyr mean), respectively. **e**, Marine sediment δ^{18}O stack[18], giving the pattern of global ice volume. **f**, Magnetic susceptibility stack record for Chinese loess[17] (normalized).

variability in the fine particle percentage (FPP)[12], which is highest during the two last glacial periods. The advection of dust to central Antarctica involves the high levels of the troposphere and the small changes in dust size may reflect changes in the altitude of transport and thus transport time[12]. Higher FPP values in glacial times have been ultimately attributed to increased isolation of Dome C during glacials, in terms of reduced dust transport associated with greater subsidence[12] or possibly through baroclinic eddies.

Comparing dust and stable isotope (δD) profiles, there is a significant correlation during glacial periods (Fig. 2), and up to 90% of the dust variability can be explained by the temperature variations. In glacial periods, most of the δD events (for example, Antarctic Isotopic Maxima) have their counterparts in the dust data shown by a reduction of dust concentrations. In contrast, dust and temperature records are not correlated during interglacial periods (Fig. 2). Indeed, the (logarithmic) relationship between dust flux and δD can
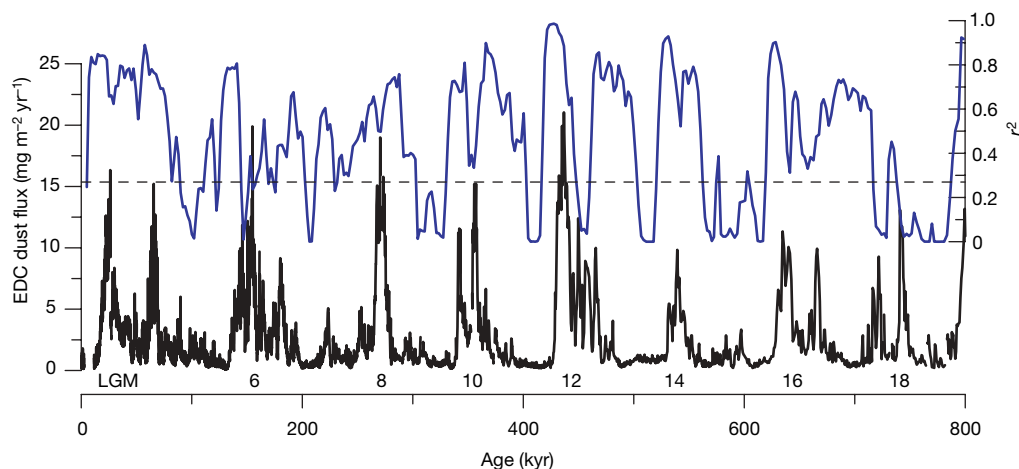
Figure 2 | EDC correlation between dust and temperature. Linear plot of dust flux (black) and the coefficient of determination r^2 (blue) between the high-pass filtered values (18-kyr cut-off) of both the δD and the logarithmic values of dust flux. The correlation was determined using 2-kyr mean values in both records and a gliding 22-kyr window. Correlations above $r^2 = 0.27$ (dashed line) are significant at a 95% confidence level. Numbers indicate the marine isotopic glacial stages.

be well fitted by a cubic polynomial (Fig. 3). Over the record, the geometric standard deviation from the polynomial fit represents a factor of ~2 in concentrations independent of the climatic period. Similarly, this relationship does not change before and after the MBE. The crescent shape of the dust–δD relationship suggests that the dust fluxes have a higher temperature sensitivity as the climate becomes colder. For δD values above about −405‰, Antarctic temperature and dust flux are not correlated, whereas there is a clear correlation for δD values below about −425‰. This behaviour may represent the expression of a progressive coupling between high- and low-latitude climate as temperatures become colder. During extreme glacial conditions the coupling appears as a direct influence of the Antarctic on the climate of southern South America[19,20]. The coupling of Antarctic and lower-latitude climate is probably coincident with the significantly extended sea ice over the Southern Atlantic and the Southern Ocean during glacials[21] and the consequent meridional (northward) shift of the atmospheric circulation (that is, the westerlies)[19–22].

Questions remain about the main factors influencing the high dust input into polar areas during glacial periods. So far, general circulation models[4,10] have reproduced a glacial dust transport and flux over tropical and mid-latitude regions which is in good agreement with global reconstructions[3], but they have failed to simulate the 25-fold increase observed in glacial dust input over Antarctica. This shortcoming is currently attributed to an incomplete representation of the source strength[23]. In addition, most of the models suggest modest changes in atmospheric transport[4,9,10], which seems supported by the relatively small changes in dust size in the EDC ice core, and by the comparison of the two EPICA ice cores[24]. Moreover, the suggestion that the 25-fold dust influx increase is mainly due to changes in source strength is challenged by evidence from South Atlantic marine records suggesting a 5- to 10-fold increase in the South American source strength[25,26] during the last glacial period.

On the basis of the new EDC data set, we suggest a new hypothesis for the glacial–interglacial changes in transport of dust. The indication (Fig. 3) that Antarctica and the southern low latitudes experienced a different coupling during the past 800 kyr is closely linked with the dust pathway within the high troposphere and the likely

extended lifetime. With respect to the dust emitted from continents, the dust arriving in Antarctica, with a mode around 2 μm diameter[27], represents the endmember of the distribution. The small size of dust particles makes *en route* gravitational settling inefficient (very long dry deposition lifetime), allowing mixing and spreading at high altitude within the troposphere. The lifetime of the particles is primarily constrained by wet deposition[23,28] and therefore by water content and temperature. As an example, along a pathway 4–6 km high and with a mean temperature of about −40 °C (conditions similar to those observed over Antarctica), a temperature reduction by 5 °C, associated with a similar change of sea surface temperature over the Southern Ocean[29], reduces the saturation water vapour pressure to about half. Under such premises, a two-dimensional model[28] obtained an increase in dust flux to Antarctica of up to a factor of 5. Thus the roughly 25-fold increase in dust flux over the Antarctic plateau during glacials could be explained by a progressive coupling of the climate of Antarctic and lower latitudes with colder temperatures, one influencing the other, and leading to the stronger aeolian deflation of southern South America and to a significantly increased dust particle lifetime along their pathway in the high-altitude troposphere over the Southern Ocean. The new EDC data set thus provides important constraints for models of the dust cycle during glacial–interglacial cycles.

METHODS SUMMARY

Apparatus. Samples for Coulter Counter Multisizer IIe measurements were obtained from discrete samples (7 cm long), decontaminated at LGGE through washing in ultrapure water. We adopted the analytical procedure described in ref. 12 (and references therein). A total of about 1,100 values have been obtained.

We obtained laser scattering data from the University of Copenhagen device for the section between 0 and 770 m. The University of Bern device was used between 770 and 3,200 m. Data from both devices were calibrated by Coulter counter. Sampling resolution is about 1 cm.

Particle size distribution. Dust size is expressed as FPP. We define FPP according to ref. 12 as the proportion of the mass of particles having diameter between 1 and 2 μm with respect to the total mass of the sample, which typically includes particles in the size range 1 to 5 μm. This parameter is inversely correlated with the modal value of the log-normal dust mass (volume) size distribution.

From a depth of 2,900 m and below, some glacial samples show distributions with an anomalously large mode. This has been attributed to particle aggregate formation in ice (Supplementary Fig. 1) and prompted us to discard all size distribution data below that point until this phenomenon is better understood. To obtain reliable dust mass, samples were submitted to ultrasonic treatment to break the aggregates apart. Between 3,139-m and 3,190-m depth, 42 samples with anomalous size distribution were submitted to ultrasonic treatment. For 39 of these we accepted the new dust mass measurement, with 11 samples showing a significantly different mass value (Supplementary Table 1). Measurement on a few chosen samples above 3,139-m depth showed normal size distribution and concentration values. However, additional measurements are scheduled for in-depth analysis of the aggregate problem.

1. Tegen, I. Modeling the mineral dust aerosol cycle in the climate system. *Quat. Sci. Rev.* **22**, 1821–1834 (2003).
2. Fung, I. *et al.* Iron supply and demand in the upper ocean. *Glob. Biogeochem. Cycles* **14**, 281–296 (2000).
3. Kohfeld, K. E. & Harrison, S. P. DIRTMAP: the geological record of dust. *Earth Sci. Rev.* **54**, 81–114 (2001).
4. Mahowald, N. *et al.* Dust sources and deposition during the last glacial maximum and current climate: A comparison of model results with paleodata from ice cores and marine sediments. *J. Geophys. Res.* **104**, 15895–15916 (1999).
5. Steffensen, J. P. The size distribution of microparticles from selected segments of the Greenland Ice Core Project ice core representing different climatic periods. *J. Geophys. Res.* **102**, 26755–26763 (1997).
6. Sun, Y. B., Clemens, S. C., An, Z. S. & Yu, Z. W. Astronomical timescale and palaeoclimatic implication of stacked 3.6-Myr monsoon records from the Chinese Loess Plateau. *Quat. Sci. Rev.* **25**, 33–48 (2006).
7. EPICA community members. Eight glacial cycles from an Antarctic ice core. *Nature* **429**, 623–628 (2004).
8. Jouzel, J. *et al.* Orbital and millennial Antarctic variability over the last 800 000 years. *Science* **317**, doi:10.1126/science.1141038 (2007).
9. Krinner, G. & Genthon, C. Tropospheric transport of continental tracers towards Antarctica under varying climatic conditions. *Tellus* **55B**, 54–70 (2003).
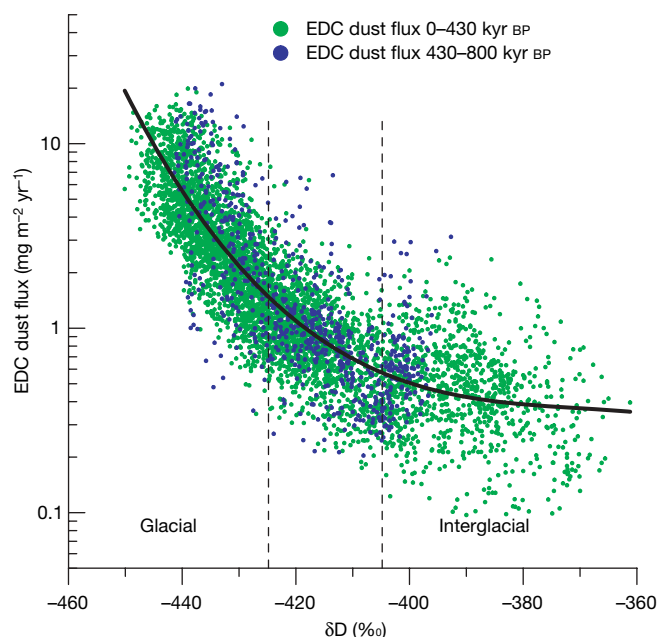

Figure 3 | EDC dust–temperature relationship. Values of δD (ref. 8) are plotted against dust flux (both at 55-cm resolution). Green and blue dots represent data from 0–430 kyr BP and 430–800 kyr BP, respectively. Superposed is a cubic polynomial fit,
$\log_{10}(f) = -3.737 \times 10^{-6}(\delta D)^3 - 4.239 \times 10^{-3}(\delta D)^2 - 1.607(\delta D) - 204$, where f is the dust flux (mg m^{-2} yr^{-1}), and δD is in ‰ ($r^2 = 0.73$, $N = 5,164$).

10. Werner, M. *et al.* Seasonal and interannual variability of the mineral dust cycle under present and glacial conditions. *J. Geophys. Res.* **107,** doi:10.1029/2002JD002365 (2002).

11. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399,** 429–436 (1999).

12. Delmonte, B. *et al.* Dust size evidence for opposite regional atmospheric circulation changes over east Antarctica during the last climatic transition. *Clim. Dyn.* **23,** 427–438 (2004).

13. Lunt, D. J. & Valdes, P. J. Dust transport to Dome C, Antarctica at the Last Glacial Maximum and present day. *Geophys. Res. Lett.* **28,** 295–298 (2001).

14. Revel-Rolland, M. *et al.* Eastern Australia: A possible source of dust in East Antarctica interglacial ice. *Earth Planet. Sci. Lett.* **249,** 1–13 (2006).

15. Wolff, E. *et al.* Southern Ocean sea-ice extent, productivity and iron flux over the past eight glacial cycles. *Nature* **440,** 491–496 (2006).

16. Fujii, Y., Kohno, M., Matoba, S., Motoyama, H. & Watanabe, O. A 320 k-year record of microparticles in the Dome Fuji, Antarctica ice core measured by laser-light scattering. *Mem. Natl Inst. Polar Res.* **57,** 46–62 (2003).

17. Kukla, G., An, Z. S., Melice, J. L., Gavin, J. & Xiao, J. L. Magnetic susceptibility record of Chinese Loess. *Trans. R. Soc. Edinb. Earth Sci.* **81,** 263–288 (1994).

18. Lisiecki, L. E. & Raymo, M. E. A Pliocene–Pleistocene stack of 57 globally distributed benthic delta O-18 records. *Paleoceanography* **20,** 1–17 (2005).

19. Iriondo, M. Patagonian dust in Antarctica. *Quat. Int.* **68,** 83–86 (2000).

20. Markgraf, V. *et al.* Paleoclimate reconstruction along the Pole—Equator—Pole transect of the Americas (PEP 1). *Quat. Sci. Rev.* **19,** 125–140 (2000).

21. Gersonde, R., Crosta, X., Abelmann, A. & Armand, L. Sea-surface temperature and sea ice distribution of the Southern Ocean at the EPILOG Last Glacial Maximum: A circum-Antarctic view based on siliceous microfossil records. *Quat. Sci. Rev.* **24,** 869–896 (2005).

22. Stuut, J.-B. W. & Lamy, F. Climate variability at the southern boundaries of the Namib (southwestern Africa) and Atacama (northern Chile) coastal deserts during the last 120,000 yr. *Quat. Res.* **62,** 301–309 (2004).

23. Mahowald, N. M. *et al.* Change in atmospheric mineral aerosol in response to climate: Last glacial period, preindustrial, modern, and doubled carbon dioxide climates. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006653 (2006).

24. Fischer, H. *et al.* Reconstruction of millennial changes in dust emission, transport and regional sea ice coverage using the deep EPICA ice cores from Atlantic and Indian Ocean sector of Antarctica. *Earth Planet. Sci. Lett.* **260,** 340–354 (2007).

25. Kumar, N. *et al.* Increased biological productivity and export production in the glacial Southern Ocean. *Nature* **378,** 675–680 (1995).

26. Chase, Z. & Anderson, R. F. Evidence from authigenic uranium for increased productivity of the glacial Subantarctic Ocean. *Paleoceanography* **16,** 468–478 (2001).

27. Delmonte, B., Petit, J.-R. & Maggi, V. Glacial to Holocene implications of the new 27000-year dust record from the EPICA Dome C (East Antarctica) ice core. *Clim. Dyn.* **18,** 647–660 (2002).

28. Yung, Y. L., Lee, T., Wang, C. H. & Shieh, Y. T. Dust: A diagnostic of the hydrologic cycle during the last glacial maximum. *Science* **271,** 962–963 (1996).

29. Stenni, B. *et al.* A late-glacial high-resolution site and source temperature record derived from the EPICA Dome C isotope records (East Antarctica). *Earth Planet. Sci. Lett.* **217,** 183–195 (2004).

LETTERS

Absolute plate motions and true polar wander in the absence of hotspot tracks

Bernhard Steinberger[1] & Trond H. Torsvik[1,2,3]

The motion of continents relative to the Earth's spin axis may be due either to rotation of the entire Earth relative to its spin axis—true polar wander[1,2]—or to the motion of individual plates[3]. In order to distinguish between these over the past 320 Myr (since the formation of the Pangaea supercontinent), we present here computations of the global average of continental motion and rotation through time[4] in a palaeomagnetic reference frame. Two components are identified: a steady northward motion and, during certain time intervals, clockwise and anticlockwise rotations, interpreted as evidence for true polar wander. We find ~18° anticlockwise rotation about 250−220 Myr ago and the same amount of clockwise rotation about 195−145 Myr ago. In both cases the rotation axis is located at about 10−20° W, 0° N, near the site that became the North American–South American–African triple junction at the break-up of Pangaea. This was followed by ~10° clockwise rotation about 145−135 Myr ago, followed again by the same amount of anticlockwise rotation about 110−100 Myr ago, with a rotation axis in both cases ~25−50° E in the reconstructed area of North Africa and Arabia. These rotation axes mark the maxima of the degree-two non-hydrostatic geoid during those time intervals, and the fact that the overall net rotation since 320 Myr ago is nearly zero is an indication of long-term stability of the degree-two geoid and related mantle structure[5,6]. We propose a new reference frame, based on palaeomagnetism, but corrected for the true polar wander identified in this study, appropriate for relating surface to deep mantle processes from 320 Myr ago until hotspot tracks can be used (about 130 Myr ago).

True polar wander (TPW) occurs because the Earth's spin axis remains always very nearly aligned with its axis of maximum non-hydrostatic moment of inertia—the short axis of the non-hydrostatic degree-two component of the geoid, which is a triaxial ellipsoid. As the geoid changes with time, the solid Earth (presumably down to the core–mantle boundary) rotates uniformly so that these axes remain aligned. Rapid TPW may occur when the short and intermediate axes of the degree-two geoid (with hydrostatic flattening removed) become nearly equal. Then the rotation axis may move along the line between those two axes, leaving the long axis at the Equator. Relating TPW to the Earth is complicated, because on geologic timescales there is presumably no rigid part that can serve as a reference frame for the entire Earth; rather, the Earth's mantle is slowly convecting, and a reference frame with zero net rotation of the convecting mantle can be regarded as most suitable. TPW has been proposed as an explanation for such disparate phenomena as rapid plate velocities, sea-level changes and the Cambrian 'Explosion of Life' (see, for example, refs 7–10). The maximum speed of TPW can be estimated on the basis of how fast the geoid changes and on how fast the Earth's hydrostatic bulge can adjust to a changing rotation axis[11]. This speed limit is estimated to be about 1° Myr^{-1} (refs 12, 13).

Modern reconstructions of the continents to give Pangaea (see, for example, ref. 14) do not look radically different from that of Wegener[3], except that continents are now placed at appropriate latitudes. However, it is not so straightforward to tie the reference frame for the Earth as a whole to observations, and hence to distinguish between absolute plate motions and TPW. Arguably the best reference points are provided by 'hotspots'—regions of intra-plate volcanism, such as in Hawaii, which are thought to be caused by upwellings from the deep mantle[15]. Chains of islands and seamounts that end at active hotspots can be explained by the motion of a rigid plate over fixed hotspots[16]. Because hotspots cannot be absolutely fixed in a convecting mantle, efforts have been made to develop a mantle reference frame in which the motion of hotspots is accounted for[17,18]. Yet their motion is presumably much slower than plate motions, and reference frames with fixed or moving hotspots are both useful in distinguishing absolute plate motions and TPW. Suitable hotspot chains exist back to about 132 Myr ago. For older times, other methods need to be used to distinguish plate motions and TPW[4,5].

Here we separate motions of continents in the palaeomagnetic reference frame into a 'mean motion' of all continents and continental motion relative to this mean. The mean motion is determined in a coordinate system illustrated in Fig. 1. West–east motion cannot
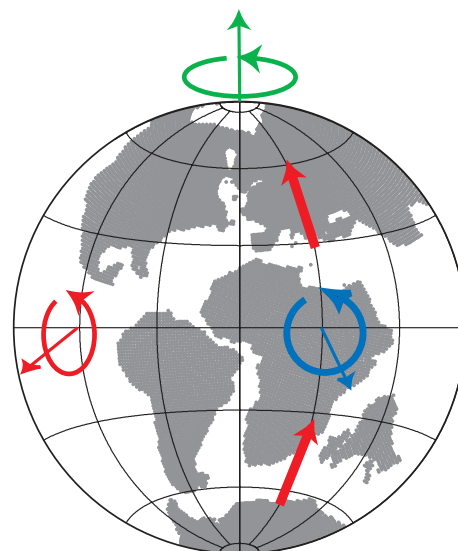
Figure 1 | Coordinate system in which the mean motion (rotation) of continents is described. One axis (blue) is on the Equator at the same longitude as the continents' centre of mass. The second axis (red) is on the Equator 90° away; rotation around this axis corresponds to coherent northward or southward motion of all continents. The third axis (green) is the Earth's spin axis; rotation around it corresponds to west–east motion.

[1]Centre for Geodynamics, Geological Survey of Norway, Leiv Eirikssons vei 39, 7491 Trondheim, Norway. [2]Physics of Geological Processes, University of Oslo, 0316 Oslo, Norway. [3]School of Geosciences, University of the Witwatersrand, Wits 2050, South Africa.

be determined in the palaeomagnetic reference frame: we additionally assume zero longitudinal motion of the African plate, which gives reasonable results with no large coherent west–east motions of all continents. Here we discuss how the characteristics of the remaining components of motion can be used in distinguishing plate motions and TPW. We restrict our analysis back to 320 Myr ago. Before that, relative motions of terranes that formed Pangaea are uncertain in relative longitude.

Figure 2 shows the cumulative mean rotation of all continents in the palaeomagnetic reference frame around an axis (blue in Fig. 1) in the equatorial plane at the same longitude as their centre of mass equivalent (black continuous line). Figure 3 shows reconstructions of continents in the same reference frame at four time intervals, and confirms that plate motions in that reference frame during these time intervals are indeed dominated by coherent rotation of all continents.

For comparison, Fig. 2 also shows corresponding rotations in two reference frames based on hotspot tracks (black dashed lines)[17–19]. Neither of them exhibits the coherent anticlockwise rotation found in the palaeomagnetic frame. The black dotted line in Fig. 2 shows the corresponding rotation in a palaeomagnetic reference frame where the coherent rotations identified above have been removed: remaining rotations are indeed much smaller.

The mean northward motion of all continents is illustrated by their cumulative mean rotation around an axis (red in Fig. 1) in the equatorial plane at a longitude 90° from their centre of mass equivalent (grey line in Fig. 2). In the palaeomagnetic reference frame, northward motion 320−185 Myr ago is followed by southward motion 185−145 Myr ago and slow overall northward motion after 145 Myr ago. After removing the coherent rotations, the mean motion of all continents becomes more evenly northward (grey dotted line, Fig. 2). This can be regarded as a continuous trend of northward motion (48° since 320 Myr ago, 24° since 240 Myr ago, 12° since 160 Myr ago) at a decreasing rate (from ∼0.3° Myr^{-1} during 320−240 Myr ago to ∼0.075° Myr^{-1} since 160 Myr ago) overlain by small fluctuations of less than 5° amplitude. For the time interval 130−120 Myr ago, this trend is also confirmed in the reference
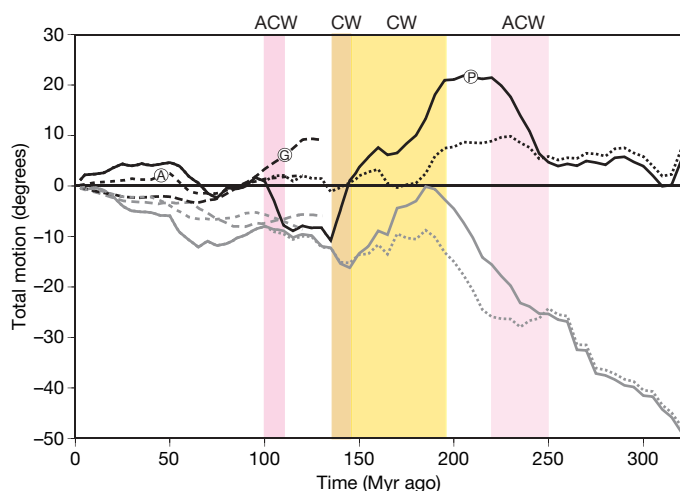
Figure 2 | Cumulative rotation and north–south motion averaged for all continents. Black lines show rotation around an equatorial axis (blue in Fig. 1) at the same longitude as the centre of mass of all continents, with positive values for a given time corresponding to clockwise (CW) rotation since that time. Grey lines show rotation around an equatorial axis (red in Fig. 1) orthogonal to the first axis, with negative values for a given time corresponding to northward motion of continents since that time. Continuous lines are computed in the palaeomagnetic reference frame (P), long dashed lines in the global mantle reference frame[17,19] (G), short dashed lines in the Africa mantle reference frame[18,19] (A), dotted lines in a new reference frame, similar to the palaeomagnetic one, except that the following rotations around an equatorial axis at constant rate have been subtracted from African plate motion: 18° anticlockwise (ACW) 250−220 Myr ago and clockwise 195−145 Myr ago around an axis at 15° W; 10° clockwise 145−135 Myr ago and anticlockwise 110−100 Myr ago around an axis at 37.5° E.

Figure 3 | Motions of continents reconstructed in the palaeomagnetic reference frame during four time intervals. Total motions are shown as black lines, connected to blue dots (locations at the beginning of the time intervals). Large blue dots with thick black lines indicate location and motion of the centre of mass of all continents. Centres of mass of African and Pacific large low-shear-velocity provinces (LLSVPs) are determined from regions with <−1% velocity anomaly in the lowermost two layers (∼300 km) of the 'smean'[27] tomography model. Eurasia is shown for reasons of simplicity as a coherent plate, but in the lowermost diagram, for example, north and south China may not have been part of Eurasia/Pangaea[14]. CAMP, central Atlantic magmatic province. (Table 1 lists the Euler rotations of South Africa in the global palaeomagnetic reference frame[19] used in this figure. The complete list in 5 Myr intervals is given in table 8 of ref. 19. These are finite rotations to move the (African) plate back from its present position to the past position at the given age. Since the longitude of a plate is not constrained palaeomagnetically, Euler pole latitude is arbitrarily set to zero.)

frames based on hotspots. There is some leeway in the location of the curve that represents the mean motion of continents (grey dotted line, Fig. 2), as there is a trade-off between the amount of northward motion and the longitude of the axes of the subtracted rotations.

Figures 2 and 3 are based on a global compilation of palaeomagnetic data and a global apparent polar wander path constructed in South African coordinates after correction for relative movements. Significant episodes of TPW should be recognizable for all locations on Earth and witnessed as rotations (derived from palaeomagnetic declinations) in the same sense on one hemisphere but with variable magnitude around an equatorially located Euler pole. As a sensitivity test for our TPW model, we calculated the rotation rates (in degrees per Myr) and sense for major plates with reasonable data coverage (Fig. 4). If TPW was much greater than 'continental drift', all plates on the same hemisphere should show the same sense of rotation, and with few exceptions we observe periods of systematic anticlockwise (110−100 Myr ago, 250−220 Myr ago) or clockwise (145−135 Myr ago, 195−145 Myr ago) rotation that conform to our TPW model. The 250−220 Myr ago phase of TPW has more speculatively been proposed[20], while the phase 110−100 Myr ago is implied in a comparison of hotspot and palaeomagnetic reference frames[19,21].

It is expected that episodes of rather 'fast' TPW (up to about $1° \, Myr^{-1}$, during time intervals when the two larger principal non-hydrostatic moments of inertia of the Earth become nearly equal) will be separated by longer time intervals of relatively slow TPW. The mean continental rotations depicted with the continuous black line in Fig. 2 and discussed above show just that characteristic, and we hence attribute their cause to be TPW. The absence of the youngest of these rotations (110−100 Myr ago) in any hotspot-based reference frame confirms our interpretation. The event of two non-hydrostatic moments of inertia becoming nearly equal does not need a specific trigger and there is hence no need to look for correlation with any other geologic events.

Under the interpretation of TPW, the axes of coherent continent rotation correspond to maxima of the degree-two non-hydrostatic geoid. For the youngest of these rotations (110–100 Myr ago), this

axis is close to the centres of the two large low-shear-velocity provinces (LLSVPs) of the lowermost mantle beneath the Pacific and Africa, ∼180° apart in longitude. The shape and extent of the LLSVPs resembles the present-day geoid highs, in particular if those features of the geoid clearly related to subduction are disregarded[22]. These axes of coherent continent rotation are 5–10° further west in a hybrid reference frame based on hotspot tracks back to 100 Myr ago and the palaeomagnetic frame before that[19], bringing the axes for the latter two time intervals (145−135 Myr ago and 110−100 Myr ago) even closer to the LLSVP centres. This proximity is a further indication of the long-term stability of LLSVPs.

The present-day long axis of the degree-two geoid does not match the LLSVP centres, as there are additional geoid highs clearly related to subduction, in particular in Indonesia and South America. During the earlier two interpreted TPW episodes, additional geoid highs due to subduction beneath the western edge of Pangaea may have caused a different location of the degree-two geoid high. For a strong increase of mantle viscosity with depth, a negative density anomaly in the lower mantle—for example, a plume head—would also cause a positive contribution to the geoid at spherical harmonic degree two[23]. The reconstructed eruption location of the central Atlantic magmatic province—potentially the largest[6] of all large igneous provinces— 200 Myr ago is almost exactly the same as the inferred rotation axis of Pangaea 250−220 Myr ago, and given approximate estimates of plume head rise times, the plume head of the central Atlantic magmatic province is likely to have been in the lower mantle 250−220 Myr ago.

Motions of continents since 130 Myr ago in a mantle reference frame have typically been rather steady. Therefore, interpreting the northward drift as motion of plates over the mantle makes sense kinematically. Furthermore, the reconstructed Pangaea location approximately overlies the African LLSVP, which is expected to correspond to a past geoid high, and therefore large motions of the Earth's spin axis towards it or away from it are not expected. Therefore, and because of the different expected characteristics of TPW (short episodes of faster motion between longer times of little motion), we regard it as unlikely that the steady northward drift represents TPW. Geometrically, a long-term northward motion of continents is linked to the preferred occurrence of spreading ridges in the Southern Hemisphere[24]. Dynamically, it may be related to the centres of the two LLSVPs of the lowermost mantle, and hence the centres of presumed large-scale mantle upwellings overlying them (beneath Africa and the South Pacific superswell), being located south of the Equator. Thus there is more upward flow in the Southern Hemisphere, and flow at shallow depth more towards the north.

The remaining fluctuations may well be due to TPW, but their magnitude of 5° or less is similar to the uncertainty of palaeomagnetic reconstructions of continents relative to the spin axis, and it is hence difficult to assess whether they are real features.

Differences between the reference frame proposed here and the palaeomagnetic frame are sufficiently small that conclusions obtained with the latter[6] should largely remain valid. The proposed

Figure 4 | Mean rotation of North America, Europe, Africa, South America and India-Pakistan since 250 Myr ago, based on palaeomagnetic declinations. We first calculated running-mean apparent polar wander paths for each plate (20 Myr window), and based on a geographic location (approximately the centre of the plate, as listed in the key) we computed mean declinations in 10 Myr intervals (bins). Declination (plate rotation) differences between successive 10 Myr bins were calculated, separated into clockwise and anticlockwise rotations, and plotted as rotation (in degrees per Myr; vertical axis). Rotation rates for each plate were then stacked. Some cases appear not to fit with the proposed TPW—for example, India-Pakistan rotating in the opposite sense 250−240 Myr ago and Africa rotating in the opposite sense 170−160 Myr ago. The latter can be explained by poor data from Africa at that time, that is, very few reliable poles and with a large scatter for the resulting mean pole (95% confidence oval of 40°). Note that some plates have near-zero rotation rates at certain intervals, and will not show up here.

Table 1 | Euler rotation parameters for South Africa

Age (Myr)	Palaeomagnetic		TPW corrected		
	Longitude (degrees)	Angle (degrees)	Latitude (degrees)	Longitude (degrees)	Angle (degrees)
100	150.8	22.3	0.0	150.8	22.3
110	169.2	32.1	3.9	153.0	26.7
135	172.6	42.1	2.1	161.0	35.9
145	155.2	38.1	-2.6	154.6	38.2
195	147.9	17.8	-2.6	155.7	35.5
220	127.2	27.2	0.3	141.6	43.1
250	145.4	40.4	4.0	145.6	40.8

Parameters are shown for palaeomagnetic[19] and TPW corrected reference frames. Note that in the former frame, latitude is zero for all ages.

TPW events would cause up to 30° latitude difference between time intervals 220−195 Myr ago and 135−100 Myr ago at locations reconstructed approximately 90° from the reconstructed location of North Africa. Results remain tentative as they are derived from continental motions only, and possible changes in relative motion between continental and oceanic plates[20] will need to be further studied in order to assess whether or not the coherent rotations found do indeed correspond to TPW. As we approach the centenary of Wegener's Pangaea[3], we are finally realizing the full spectrum of available palaeogeographic and kinematic tools, integrated with a fully dynamic global framework.

METHODS SUMMARY

Relative plate motions since 320 Myr ago have been compiled by Torsvik *et al.*[19]. For absolute plate motions, we use their palaeomagnetic reference frame[19]. After 130 Myr ago, we also use the 'global mantle'[17,19] and 'Africa mantle'[18,19] reference frames for comparison. For each plate, we only consider the 'continental' part, with elevation > -200 m. From the present-day positions and past motions of continents, their positions through time are determined. From these continent distributions and motions, the quantities equivalent to centre of mass, inertia tensor I and angular momentum $\mathbf{L}$, of all continents, but with mass replaced by area, relative to the centre of the Earth, are computed. By inverting the relation $\mathbf{L} = I\boldsymbol{\omega}$, the mean rotation $\boldsymbol{\omega}$ of all continents is determined. At each time, three components are determined in an orthogonal coordinate system, with one axis along the Earth's spin axis, and one at the same longitude as the equivalent to the centre of mass of all continents. These components are integrated through time, yielding cumulative rotations. These approximately represent finite rotations, but not exactly, as the centre of mass longitude also changes with time in our reference frame, although not by much. In both the palaeomagnetic reference frame and the modified reference frame proposed here, the cumulative rotation around the Earth's spin axis (corresponding to mean west–east motion of all continents, which is not determined palaeomagnetically) is very small over the period 320−130 Myr ago (less than 5° fluctuations), which appears reasonable.

Figures were prepared using the GMT[25] and GMAP[26] software.

1. Gold, T. Instability of the Earth's axis of rotation. *Nature* 175, 526–529 (1955).
2. Goldreich, P. & Toomre, A. Some remarks on polar wandering. *J. Geophys. Res.* 74, 2555–2569 (1969).
3. Wegener, A. The origin of the continents. *J. Geodyn.* 32, 29–63 (2001); Die Entstehung der Kontinente. *Petermanns Geographische Mitteilungen* 58(I), 185–195; 253–256; 305–309 (1912).
4. Jurdy, D. M. & Van Der Voo, R. True polar wander since the Early Cretaceous. *Science* 187, 1193–1196 (1975).
5. Evans, D. A. D. True polar wander and supercontinents. *Tectonophysics* 362, 303–320 (2003).
6. Torsvik, T. H., Smethurst, M. A., Burke, K. & Steinberger, B. Large igneous provinces generated from the margins of the large low-velocity provinces in the deep mantle. *Geophys. J. Int.* 167, 1447–1460 (2006).
7. Kirschvink, J. L., Ripperdan, R. L. & Evans, D. A. Evidence for a large-scale reorganization of Early Cambrian continental landmasses by inertial interchange true polar wander. *Science* 277, 541–545 (1997).
8. Mound, J. E., Mitrovica, J. X. & Milne, G. A. Sea-level change and true polar wander during the Late Cretaceous. *Geophys. Res. Lett.* 28, 2057–2060 (2001).
9. Torsvik, T. H., Van der Voo, R. & Redfield, T. F. Relative hotspot motions versus true polar wander. *Earth Planet. Sci. Lett.* 202, 185–200 (2002).
10. Maloof, A. C. et al. Combined paleomagnetic, isotopic, and stratigraphic evidence for true polar wander from the Neoproterozoic Akademikerbreen Group, Svalbard. *Geol. Soc. Am. Bull.* 118, 1099–1124 (2006).
11. Steinberger, B. & O'Connell, R. J. Changes of the Earth's rotation axis owing to advection of mantle density heterogeneities. *Nature* 387, 169–173 (1997).
12. Steinberger, B. & O'Connell, R. J. in *Ice Sheets, Sea Level and the Dynamic Earth* (eds Mitrovica, J. X. & Vermeersen, L. L. A.) 233–256 (Geodynamics Series vol. 29, AGU, Washington DC, 2002).
13. Tsai, V. C. & Stevenson, D. J. Theoretical constraints on true polar wander. *J. Geophys. Res.* 112, B05415, doi:10.1029/2005JB003923 (2007).
14. Torsvik, T. H. & Cocks, L. R. M. Earth geography from 400 to 250 million years: A palaeomagnetic, faunal and facies review. *J. Geol. Soc. Lond.* 161, 555–572 (2004).
15. Wilson, J. T. A possible origin of the Hawaiian Islands. *Can. J. Phys.* 41, 863–870 (1963).
16. Morgan, W. J. Convection plumes in the lower mantle. *Nature* 230, 42–43 (1971).
17. Steinberger, B., Sutherland, R. & O'Connell, R. J. Prediction of Emperor-Hawaii seamount locations from a revised model of global plate motion and mantle flow. *Nature* 430, 167–173 (2004).
18. O'Neill, C., Müller, D. & Steinberger, B. On the uncertainties in hotspot reconstructions, and the significance of moving hotspot reference frames. *Geochem. Geophys. Geosyst.* 6, Q04003, doi:10.1029/2004GC000784 (2005).
19. Torsvik, T. H., Müller, R. D., Van der Voo, R., Steinberger, B. & Gaina, C. Global plate motion frames: Toward a unified model. *Rev. Geophys.* (in the press); preprint at ⟨http://www.geodynamics.no/guest/GlobalRef.pdf⟩.
20. Marcano, M. C., Van der Voo, R. & Mac Niocaill, C. True polar wander during the Permo-Triassic. *J. Geodyn.* 28, 75–95 (1999).
21. Besse, J. & Courtillot, V. Apparent and true polar wander and the geometry of the geomagnetic field over the last 200 Myr. *J. Geophys. Res.* 107, 2300, doi:10.1029/2000JB000050 (2002).
22. Hager, B. H. Subducted slabs and the geoid: Constraints on mantle rheology and flow. *J. Geophys. Res.* 89, 6003–6015 (1984).
23. Richards, M. A. & Hager, B. H. Geoid anomalies in a dynamic Earth. *J. Geophys. Res.* 89, 5987–6002 (1984).
24. McCarthy, D. Geophysical explanation for the disparity in spreading rates between the Northern and Southern hemispheres. *J. Geophys. Res.* 112, B03410, doi:10.1029/2006JB004535 (2007).
25. Wessel, P. & Smith, W. H. F. Free software helps map and display data. *Eos* 72, 441 (1991).
26. Torsvik, T. H. & Smethurst, M. A. Plate tectonic modeling: Virtual reality with GMAP. *Comput. Geosci.* 25, 395–402 (1999).
27. Becker, T. W. & Boschi, L. A comparison of tomographic and geodynamic mantle models. *Geochem. Geophys. Geosyst.* 3, 1003, doi:10.1029/2001GC000168 (2002).

nature

METHODS

We express Euler rotations for times t_A as 3×3 rotation matrices[28] A, that is, a point on the plate at present at location $\mathbf{x}$ (in cartesian coordinates) was at $A\mathbf{x}$ at time t_A. Points on other plates are reconstructed by combining rotations of plates relative to each other (compiled in table 1 of ref. 19) with the Euler rotations of South Africa (Table 1).

Maps in Fig. 3 are computed by rotating all points on a one-degree grid with elevation $> -200\,m$ from their present location to the location at that time. Present-day NUVEL[29] plate boundaries are used to assign points to plates. This procedure is obviously not exactly correct—for example, Greenland has not always been part of the North American plate. However, this simplification affects our results only in a minor way.

In order to obtain the TPW corrected rotations, we first compute stage rotations, that is, finite rotations for 5 Myr time intervals. With the Euler rotations for times t_A and t_B expressed as rotation matrices A and B, the stage rotation to move the plate from its position at t_A back to its position at t_B is then $S_{AB} = BA^{-1}$. Stage rotations are again expressed as rotation vectors. TPW corrections, which are derived by visual inspection of Figs 2 and 3, and specified in Fig. 2 legend (description of dotted line), are then applied by vector subtraction in cartesian coordinates. TPW corrected stage rotations for time intervals $t_B - t_A$ are again expressed as rotation matrices[28] $S_{AB,corr}$. The TPW corrected rotation matrix for t_B is then $B_{corr} = S_{AB,corr}A$, and corrected finite rotations are computed successively backward in time. We note that, even though TPW corrections for the time 145 Myr ago, and for times 250 Myr ago and before, are not intended, our procedure also yields slightly different TPW corrected Euler rotations for these times. However, differences are much smaller than for those times where the corrections are indeed intended.

Dotted lines in Fig. 2 are computed using the TPW corrected reference frame in the same way as continuous lines were computed with the palaeomagnetic reference frame.

Torsvik *et al.*[6] concluded that reconstructed large igneous province (LIP) eruption sites overlay the edges of LLSVPs, using four different reference frames after 132 Myr ago and the palaeomagnetic frame before that. The change of reference frame before 132 Myr ago only possibly affects the central Atlantic magmatic province (CAMP) at 200 Myr ago and the Karroo LIP at 182 Myr ago. As explained, the reconstructed CAMP location coincides with the axis of added rotation and hence remains unaffected; Karroo is rotated by about $8°$, leaving it close to the LLSVP margin.

28. Chang, T., Stock, J. & Molnar, P. The rotation group in plate tectonics and the representation of uncertainties of plate reconstructions. *Geophys. J. Int.* **101**, 649–661 (1990).
29. DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Current plate motions. *Geophys. J. Int.* **101**, 425–478 (1990).

LETTERS

Localization and functionality of microsporidian iron–sulphur cluster assembly proteins

Alina V. Goldberg[1]*, Sabine Molik[2]*, Anastasios D. Tsaousis[3], Karina Neumann[2], Grit Kuhnke[2], Frederic Delbac[4], Christian P. Vivares[4], Robert P. Hirt[1], Roland Lill[2] & T. Martin Embley[1]

Microsporidia are highly specialized obligate intracellular parasites of other eukaryotes (including humans[1]) that show extreme reduction at the molecular, cellular and biochemical level[2,3]. Although microsporidia have long been considered as early branching eukaryotes that lack mitochondria[4], they have recently been shown to contain a tiny mitochondrial remnant called a mitosome[2,5]. The function of the mitosome is unknown, because microsporidians lack the genes for canonical mitochondrial functions, such as aerobic respiration and haem biosynthesis. However, microsporidial genomes encode several components of the mitochondrial iron–sulphur (Fe–S) cluster assembly machinery. Here we provide experimental insights into the metabolic function and localization of these proteins. We cloned, functionally characterized and localized homologues of several central mitochondrial Fe–S cluster assembly components for the microsporidians _Encephalitozoon cuniculi_ and _Trachipleistophora hominis_. Several microsporidial proteins can functionally replace their yeast counterparts in Fe–S protein biogenesis. In _E. cuniculi_, the iron (frataxin) and sulphur (cysteine desulphurase, Nfs1) donors and the scaffold protein (Isu1) co-localize with mitochondrial Hsp70 to the mitosome, consistent with it being the functional site for Fe–S cluster biosynthesis. In _T. hominis_, mitochondrial Hsp70 and the essential sulphur donor (Nfs1) are still in the mitosome, but surprisingly the main pools of Isu1 and frataxin are cytosolic, creating a conundrum of how these key components of Fe–S cluster biosynthesis coordinate their function. Together, our studies identify the essential biosynthetic process of Fe–S protein assembly as a key function of microsporidian mitosomes.

The biosynthesis of Fe–S clusters is the only known essential biosynthetic function of yeast mitochondria[6]. It is required for the maturation of mitochondrial and cytosolic proteins of diverse functions, and, critically, for the generation of a functional cellular protein synthesis machinery[7]. Fe–S protein biogenesis by the _Saccharomyces cerevisiae_ mitochondrial Fe–S cluster assembly machinery involves the _de novo_ synthesis of a transiently bound Fe–S cluster on the scaffold proteins Isu1 and Isu2. This step requires the cysteine desulphurase complex Nfs1–Isd11 (ref. 8) as a sulphur donor, frataxin (Yfh1) as a putative iron donor and the input of electrons from ferredoxin (Yah1). Fe–S cluster transfer from Isu1 and Isu2 to target apoproteins is facilitated by a specialized mitochondrial Hsp70 (Ssq1), its co-chaperone Jac1 and the monothiol glutaredoxin Grx5 (ref. 9). Phylogenetic analyses of Nfs1 (ref. 10) and Isu1 (ref. 11) suggest that they were inherited from the mitochondrial endosymbiont, making Fe–S cluster assembly potentially one of the aboriginal contributions of the protomitochondrion to the eukaryotic cell.

We cloned several _E. cuniculi_ and _T. hominis_ homologues of the yeast mitochondrial Fe–S cluster assembly machinery including Isu1, Isd11, Nfs1, Yfh1, Ssq1 and Grx5. Sequence alignments showed that key amino acid residues required for the activity of yeast, human and bacterial Fe–S cluster proteins are conserved in these proteins (Supplementary Figs 1–6). Proteins of the expected size were detected in western blots of _T. hominis_ extracts using specific antisera (Supplementary Fig. 7). We tested the functionality of the _E. cuniculi_ and _T. hominis_ Fe–S cluster proteins by complementation of yeast mutants deficient in the relevant components of mitochondrial Fe–S cluster biosynthesis[12]. As we consistently failed to detect any targeting signals in the microsporidial Fe–S cluster proteins using the available prediction methods[13], efficient localization to the yeast mitochondrion was ensured by fusion of a well characterized fungal mitochondrial presequence to their amino termini[14].

Transformation of a yeast _ISU_ mutant cell (strain Gal-ISU1/Δisu2 in which the _ISU2_ gene is deleted and the _ISU1_ gene can be repressed by growth in the presence of glucose[14]) with a plasmid encoding _T. hominis_ (Th)Isu1 fully rescued the growth defect observed after depletion of yeast Isu1 (Fig. 1a). Similarly, ThIsu1 supported the maturation of Fe–S proteins in yeast as measured by the activity of the mitochondrial Fe–S enzymes aconitase and succinate dehydrogenase (Fig. 1b), or by the _de novo_ maturation of aconitase and the cytosolic Fe–S protein isopropylmalate isomerase (Leu1) in a radiolabelling assay[14] using ⁵⁵Fe (Fig. 1c). The degree of complementation by ThIsu1 was similar to that by extrachromosomal expression of yeast Isu1. An additional measure of the ability of ThIsu1 to support cellular Fe–S protein assembly was provided by the observation that the mitochondrial iron accumulation developing after functional defects of the Fe–S cluster machinery[14] was reduced in the Gal-ISU1/Δisu2 yeast mutant producing ThIsu1 (Fig. 1d). These data clearly demonstrate that ThIsu1 can replace the endogenous Isu proteins as functional scaffold proteins in cellular Fe–S protein biosynthesis, and hence is a functional homologue of yeast Isu1.

No complementation was observed on expression of ThNfs1 in Nfs1-depleted Gal-NFS1 yeast cells (data not shown). When ThNfs1 was produced in _Escherichia coli_, most of the protein was found in inclusion bodies (Fig. 1e and data not shown) and we failed to detect ThNfs1-specific cysteine desulphurase activity in any fraction. It has recently been shown[8] that yeast Nfs1 is stabilized by complex formation with the small protein Isd11. We therefore co-expressed ThNfs1 and ThIsd11, the microsporidian homologue of yeast Isd11. In this experiment, both ThNfs1 and ThIsd11 were co-isolated with _E. coli_ membranes that contained ThNfs1- and ThIsd11-specific cysteine desulphurase activity (Fig. 1f). Our data suggest that ThIsd11 stabilizes

ThNfs1 by complex formation and, as in yeast[8], that this complex represents the functional cysteine desulphurase of *T. hominis*.

Targeting of the *E. cuniculi* proteins EcYfh1 and EcGrx5 to mitochondria rescued growth of the respective regulatable yeast Fe–S cluster mutants (Supplementary Fig. 8), and restored both the activities (Fig. 2a) and the *de novo* assembly of mitochondrial Fe–S proteins (Fig. 2b). Notably, production of EcYfh1 without a mitochondrial presequence did not restore growth and function, but led to efficient degradation of the protein (Supplementary Fig. 8c and data not shown). These data identify EcYfh1 and EcGrx5 as functional homologues of their yeast counterparts. No rescue of the phenotypes of the respective yeast mutants was observed for *EcISU1* and *EcNFS1* genes from *E. cuniculi* (data not shown). Apparently, these proteins do not functionally replace the homologous yeast Fe–S cluster components under the assay conditions.

On expression of ThmHsp70 (a mitosomal marker protein of *T. hominis*), a further decrease, rather than an increase, in the reduced growth rate of Ssq1-depleted Gal-SSQ1 cells[15] was observed (data not shown). Even in wild-type yeast cells overproduction of ThmHsp70 severely diminished the growth rate (Supplementary Fig. 9a). To analyse whether this decreased growth rate was due to a specific dominant-negative effect of heterologous ThmHsp70 on the

function of the yeast mitochondrial Fe–S cluster assembly machinery, we measured Fe–S protein maturation. Synthesis of ThmHsp70 in yeast significantly decreased the *de novo* assembly of mitochondrial and cytosolic Fe–S proteins (Supplementary Fig. 9b). To define further the reason for this effect, we tested whether overproduction of ThmHsp70 leads to an accumulation of Fe–S clusters on the scaffold protein Isu1. Ssq1 is required for Fe–S cluster transfer from Isu1 to recipient apoproteins, and consequently its functional impairment leads to an accumulation of Fe–S clusters on Isu1 (refs 15, 16). After overproduction of ThmHsp70 we observed a more than twofold increase of radiolabelled ^{55}Fe–S clusters on Isu1 (Supplementary Fig. 9c). These antipodal effects of ThmHsp70 on Fe–S cluster association on Fe–S target proteins and Isu1 are similar to previous observations on depletion of yeast Ssq1 (ref. 15). They suggest that the transfer of Fe–S clusters from Isu1 to apoproteins is inhibited by ThmHsp70, and that ThmHsp70 specifically interferes with the normal function of yeast Ssq1.

These functional data raised the question as to the subcellular localization of the microsporidial proteins. To determine the location of the critical components Isu1, Nfs1 and frataxin (Yfh1) in both *E. cuniculi* and *T. hominis*, we used indirect immunofluorescence analyses (IFA) of parasite-infected rabbit kidney cells[5], using antisera
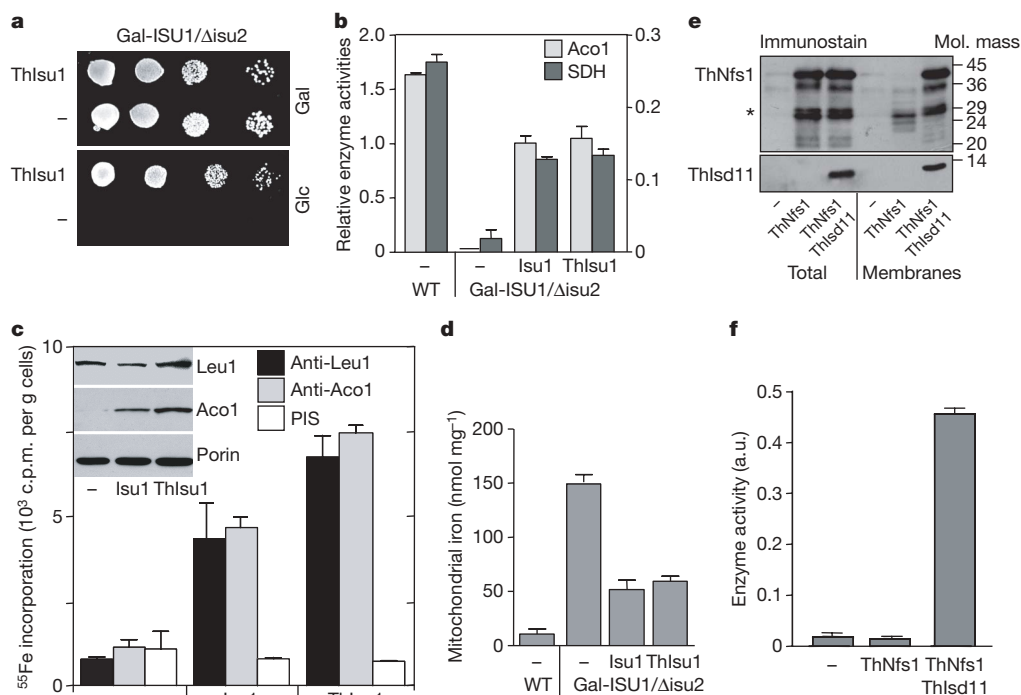
Figure 1 | *Trachipleistophora hominis* Fe–S cluster proteins are functional in Fe–S cluster biosynthesis reactions. a, A chromosomal mutant Gal-ISU1/Δisu2 lacks the *ISU2* gene yet allows the expression of *ISU1* under the control of the *GAL1-10* promoter, which is induced in the presence of galactose (Gal) and repressed by glucose (Glc), leading to a severe growth defect[14]. Expression of ThIsu1 carrying a mitochondrial targeting sequence from *Neurospora crassa* F_1-ATPase subunit β fully restores wild-type growth on glucose. **b,** Synthesis of mitochondria-targeted ThIsu1 in Isu1-depleted cells supports mitochondrial Fe–S protein assembly. Gal-ISU1/Δisu2 cells were transformed with plasmid p426 containing either no insert (−) or genes encoding ThIsu1 or yeast Isu1. Mitochondria were purified, the activities of two mitochondrial Fe–S cluster-containing enzymes, aconitase (Aco1) and succinate dehydrogenase (SDH), were measured and plotted relative to the activity of a non-Fe–S-containing mitochondrial enzyme, malate dehydrogenase (MDH). A control experiment was performed with wild-type (WT) cells. Error bars indicate s.e.m. (*n* = 3). **c,** Mitochondria-targeted ThIsu1 supports *de novo* Fe–S protein assembly of both aconitase and cytosolic isopropylmalate isomerase (Leu1). Gal-ISU1/Δisu2 cells, transformed with plasmid p426 containing either no insert (−) or genes encoding yeast Isu1 or ThIsu1, were radiolabelled with ^{55}Fe, and a cell

extract was prepared[12]. Immunoprecipitation using anti-Leu1 or anti-aconitase antibodies or pre-immune serum (PIS) was performed, and co-precipitated ^{55}Fe was quantified by scintillation counting. The inset shows western blots of Leu1, Aco1 and the loading control porin from samples used for ^{55}Fe incorporation assays. In the absence of functional Isu proteins, the apo form of Aco1 is highly sensitive to degradation and was present in low amounts only. Error bars indicate s.e.m. (*n* = 4). c.p.m., counts per min. **d,** Expression of mitochondria-targeted ThIsu1 reduces the accumulation of free iron in yeast mitochondria depleted of their Isu proteins. Non-haem, non-Fe–S ('free') iron associated with isolated mitochondria from **b** was measured by the bathophenantroline method[12]. Error bars indicate s.e.m. (*n* = 3). **e,** ThNfs1 and ThIsd11 carrying an N-terminal His tag were co-expressed in *E. coli*. Cell extracts (Total) were used for the isolation of the membrane fraction by ultracentrifugation. Aliquots were analysed by SDS–polyacrylamide gel electrophoresis and immunostaining with antibodies against ThNfs1 or the His tag of ThIsd11. The asterisk indicates degradation products of ThNfs1. **f,** Membranes as isolated in **e** were lysed in buffer containing 1.5% digitonin, and the cysteine desulphurase activity was estimated (a.u., arbitrary units)[27]. Error bars indicate s.e.m. (*n* = 3).

raised against recombinant *T. hominis* proteins or to recombinant EcmHsp70, as a marker for the *E. cuniculi* mitosome. In *E. cuniculi*, EcIsu1, EcNfs1 and EcYfh1 co-localized with mitosomal EcmHsp70 (Fig. 3). These data demonstrate that core elements of the Fe–S cluster biosynthetic machinery localize to the *E. cuniculi* mitosome, consistent with the model developed for yeast mitochondria[9]. By contrast, in *T. hominis*, the cysteine desulphurase ThNfs1 and ThmHsp70 (ref. 5) co-localized to the mitosome, but the major pools of ThIsu1 and ThYfh1 are, so far uniquely for a eukaryote, in the

Figure 2 | Mitochondrial-targeted EcYfh1 and EcGrx5 (partially) restore Fe–S protein maturation in Yfh1- and Grx5-depleted yeast cells. a, Mitochondria were isolated from wild-type cells (WT), Gal-YFH1 cells lacking (−) or expressing the fusion gene *pSu9-EcYFH1*, and Gal-GRX5 cells lacking (−) or expressing the fusion gene *pSu9-EcGRX5* after growth in minimal medium containing glucose to deplete Yfh1 or Grx5 (see Supplementary Fig. 8). The enzyme activities of aconitase and succinate dehydrogenase (SDH) were measured and plotted relative to the activities of malate dehydrogenase (MDH). Error bars indicate s.e.m. (*n* = 4). **b,** EcYfh1 and EcGrx5 support the d*e novo* assembly of mitochondrial Fe–S proteins. Biotin synthase (Bio2) was overproduced in cells described in **a**, and cells were grown in iron-poor minimal medium containing glucose. Fe–S protein assembly was measured by the radiolabelling and immunoprecipitation assay described in Fig. 1c using antibodies specific for biotin synthase. Synthesis of mitochondria-targeted EcYfh1 or EcGrx5 restored ^{55}Fe–S cluster incorporation into Bio2. Error bars indicate s.e.m. (*n* = 4).

cytosol (Fig. 4). We cannot exclude by our IFA data the possibility that there is a minor pool of ThIsu1 and ThYfh1 in the *T. hominis* mitosome, or of ThNfs1 in the cytosol. This disjunct distribution raises intriguing questions as to how the separated Fe–S cluster components can perform their usually tightly coordinated function. The functional data already demonstrate that ThIsu1 accepts sulphur from *Saccharomyces* mitochondrial Nfs1, because it can complement the yeast Isu1 and Isu2 mitochondrial mutant (Fig. 1). How the persulphide moiety generated by the cysteine desulphurase activity of mitosomal ThNfs1 might be delivered to a cytosolic ThIsu1 scaffold is currently unknown. However, even in well studied model systems like yeast[9] there are still uncertainties about what essential product of mitochondrial Fe–S biosynthesis is actually exported from mitochondria to support cytosolic Fe–S biosynthesis[17].

Until recently it was thought that some eukaryotes including *Giardia lamblia*, microsporidians like *E. cuniculi* and *T. hominis*, and *Trichomonas vaginalis* were descended from anaerobic nucleated cells that split from other eukaryotes before the mitochondrial symbiosis[4]. All of these species were subsequently shown to possess double-membraned organelles, called mitosomes—or hydrogenosomes when they make hydrogen—that import proteins using mitochondria-like protein import pathways[4]. These discoveries underpin the current paradigm that all eukaryotes contain an organelle of mitochondrial ancestry: a mitochondrial homologue. A testable prediction of this hypothesis is that mitochondrial homologues should have maintained at least one vital role that makes them indispensable for eukaryotic cells. Comparison of the genomes of microbial eukaryotes with mitosomes[18,19] and hydrogenosomes[20] has revealed that very little of the 'metabolome' of yeast and human mitochondria is conserved across the eukaryotic tree. One pathway that is conserved, in one form or another[21], is that for the biosynthesis of Fe–S clusters for insertion into Fe–S proteins—ancient and essential components of eukaryotic and prokaryotic cellular metabolism[7,22]. Key proteins for the Fe–S pathway occur in human and plant mitochondria[9,23,24] and in the *Giardia* mitosome[11] and *Trichomonas* hydrogenosome[25]. This remarkable degree of conservation underpins the hypothesis that the biosynthesis of Fe–S clusters might be a central function of mitochondrial homologues[6,7,9].

Microsporidia provide an important test of this hypothesis because unlike *Giardia* and *Trichomonas,* which are extracellular anaerobic mucosal parasites, microsporidia are derived fungi that have undergone extreme genomic and cellular reduction as specialized obligate intracellular parasites of aerobic eukaryotes[1,2]. The mitosomes of microsporidia are tiny (~50 nm by 90 nm) and structurally undifferentiated[5,26], they lack a genome and their protein import pathway is rudimentary[3]. In addition, there are no obvious ways for the organelle to make or import ATP (ref. 2). Nevertheless, the genome of *E. cuniculi* still encodes the core elements of the mitochondrial Fe–S cluster machinery[2], which we have now located experimentally to its mitosomes. *Trachipleistophora hominis* has relocated the major pools of Isu1 and frataxin to the cytosol, but its mitosome contains at least two essential components, Nfs1 and mHsp70, of this pathway. These data strongly suggest that a key role of the mitosome is in the biosynthesis of microsporidian cytosolic Fe–S proteins, including Nar1 and Rli1 (ref. 2), which are vital for cell survival[9].

METHODS SUMMARY

Genes for various *E. cuniculi* and *T. hominis* Fe–S cluster proteins were cloned into a yeast vector encoding a mitochondrial-targeting signal. Vectors were transformed into regulatable yeast mutants in which different proteins of the mitochondrial Fe–S cluster pathway can be depleted. Functional heterologous complementation was measured by either growth in media containing the appropriate carbon sources, radiolabelling experiments of Fe–S proteins using ^{55}Fe, or the measurement of mitochondrial and cytosolic enzyme activities. For co-expression of *T. hominis* cysteine desulphurase (ThNfs1) and ThIsd11 in *E. coli*, we used the pET-DUET1 vector. After expression, a cell extract was produced, membranes were isolated by centrifugation and used for assaying cysteine
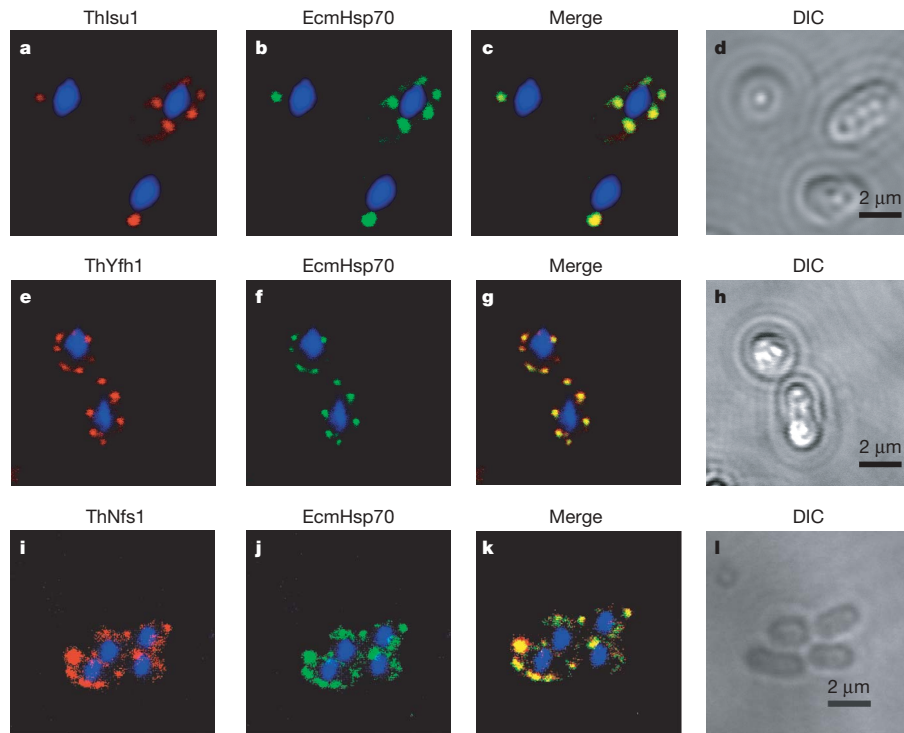
Figure 3 | Cellular localization of Isu1-, Yfh1- and Nfs1-like proteins in *E. cuniculi* by immunofluorescence and confocal microscopy. The nuclei of purified parasites were labelled with 4,6-diamidino-2-phenylindole (DAPI; blue). **a–c**, Rabbit antisera (1:10 dilution) to *T. hominis* ThIsu1 (**a**, red) co-localizes with biotinylated rabbit antisera (1:10) raised to the mitosomal marker protein EcmHsp70, labelling (**b**, green; **c**, yellow merge) discrete structures reminiscent of *T. hominis*[5] mitosomes (see also Fig. 4). **e–g**, Rabbit antisera (1:10) to ThYfh1 (**e**, red) co-localizes with EcmHsp70 (**f**, green; **g**, yellow merge). **i–k**, Rabbit antisera (1:50) to ThNfs1 peptides (**i**, red) co-localizes with EcmHsp70-positive structures (**j**, green; **k**, yellow merge). **d, h, l**, Differential interference contrast (DIC) images of the cells used for immunofluorescence.
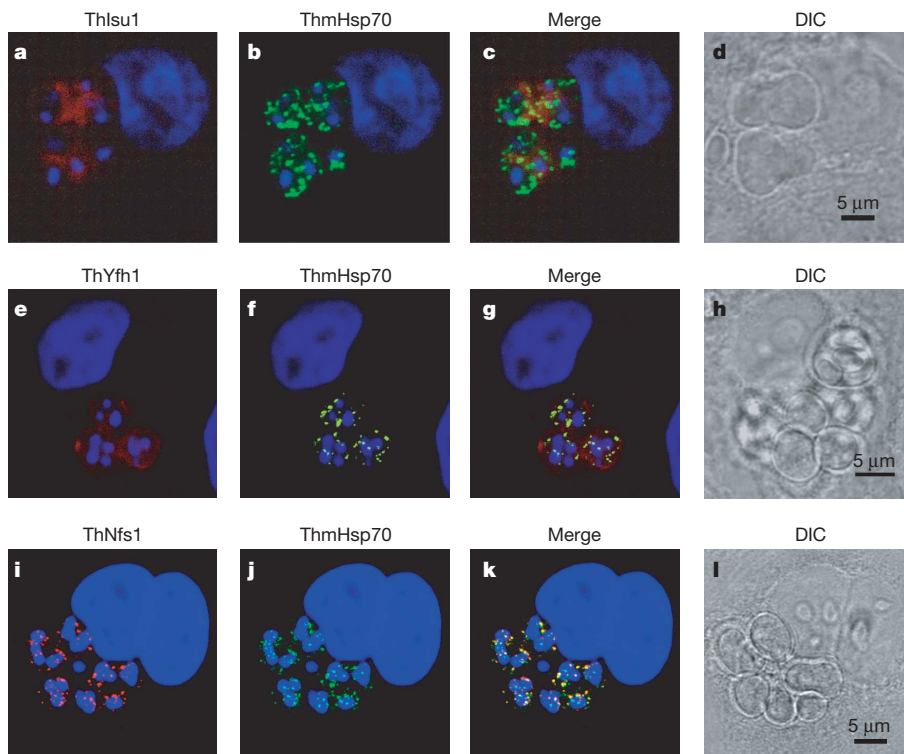
Figure 4 | Cellular localization of Isu1-, Yfh1- and Nfs1-like proteins in *T. hominis*. The nuclei of rabbit kidney host cells (large nuclei) and parasites (small nuclei) were labelled with DAPI (blue). **a–c**, Mouse antisera (1:10 dilution) to ThIsu1 detects the cytosol (**a**, red) of *T. hominis*, whereas the rabbit antibody (1:500) raised to the mitosomal marker protein ThmHsp70 detects (**b**, green; **c**, merge) discrete structures corresponding to mitosomes[5]. **e–g**, Rat antisera (1:10) to ThYfh1 detects the cytosol (**e**, red) of *T. hominis*, whereas ThmHsp70 locates to mitosomes (**f**, green; **g**, merge). **i–k**, Rabbit antisera (1:50) to ThNfs1 peptides (**i**, red) co-localizes with discrete structures labelled with biotinylated antibody (1:10) to ThmHsp70 (**j**, green; **k**, yellow merge). **d, h, l**, DIC images of the cells used for immunofluorescence.

desulphurase activity. For antibody production, *T. hominis ThISU1*, *ThNFS1* and *ThYFH1* were cloned into the pQE-40 expression vector and expressed in *E. coli*. Antibodies against ThNfs1 were also generated against synthetic peptides comprising the most variable domains of the protein. Expressed proteins were purified by gel electrophoresis for the generation of mouse, rabbit or rat antisera. For indirect fluorescence immunolocalization (IFA) studies, RK-13 cells infected with either *E. cuniculi* or *T. hominis* were grown and fixed on coverslips before further processing for IFA. Cells of *E. cuniculi* were also purified by disruption of infected RK-13 cells followed by filtration through glass wool and fixation on glass slides. After blocking, slides were incubated with the relevant antisera raised for each *T. hominis* protein and subsequently incubated with fluorescent dye-conjugated secondary antibodies: Alexa 633 (red), Alexa 594 (red), or Alexa 488 (green). In IFA experiments where both antisera in an assay had been raised in rabbit, we biotinylated one of the antibodies to enable detection via a streptavidin conjugate. The labelled cells were visualized under a laser scanning confocal microscope.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Keeling, P. J. & Fast, N. M. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.* **56**, 93–116 (2002).
2. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi. Nature* **414**, 450–453 (2001).
3. Burri, L., Williams, B. A., Bursac, D., Lithgow, T. & Keeling, P. J. Microsporidian mitosomes retain elements of the general mitochondrial targeting system. *Proc. Natl Acad. Sci. USA* **103**, 15916–15920 (2006).
4. Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623–630 (2006).
5. Williams, B. A., Hirt, R. P., Lucocq, J. M. & Embley, T. M. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis. Nature* **418**, 865–869 (2002).
6. Lill, R. & Kispal, G. Maturation of cellular Fe-S proteins: an essential function of mitochondria. *Trends Biochem. Sci.* **25**, 352–356 (2000).
7. Kispal, G. *et al.* Biogenesis of cytosolic ribosomes requires the essential iron-sulphur protein Rli1p and mitochondria. *EMBO J.* **24**, 589–598 (2005).
8. Wiedemann, N. *et al.* Essential role of Isd11 in mitochondrial iron-sulfur cluster synthesis on Isu scaffold proteins. *EMBO J.* **25**, 184–195 (2006).
9. Lill, R. & Muhlenhoff, U. Iron-sulfur protein biogenesis in eukaryotes: components and mechanisms. *Annu. Rev. Cell Dev. Biol.* **22**, 457–486 (2006).
10. Emelyanov, V. V. Phylogenetic affinity of a *Giardia lamblia* cysteine desulfurase conforms to canonical pattern of mitochondrial ancestry. *FEMS Microbiol. Lett.* **226**, 257–266 (2003).
11. Tovar, J. *et al.* Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature* **426**, 172–176 (2003).
12. Molik, S., Lill, R. & Muhlenhoff, U. Methods for studying iron metabolism in yeast mitochondria. *Methods Cell Biol.* **80**, 261–280 (2007).
13. Guda, C., Fahy, E. & Subramaniam, S. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* **20**, 1785–1794 (2004).
14. Gerber, J., Neumann, K., Prohl, C., Muhlenhoff, U. & Lill, R. The yeast scaffold proteins Isu1p and Isu2p are required inside mitochondria for maturation of cytosolic Fe–S proteins. *Mol. Cell. Biol.* **24**, 4848–4857 (2004).
15. Muhlenhoff, U., Gerber, J., Richhardt, N. & Lill, R. Components involved in assembly and dislocation of iron-sulfur clusters on the scaffold protein Isu1p. *EMBO J.* **22**, 4815–4825 (2003).
16. Dutkiewicz, R. *et al.* The Hsp70 chaperone Ssq1p is dispensable for iron-sulfur cluster formation on the scaffold protein Isu1p. *J. Biol. Chem.* **281**, 7801–7808 (2006).
17. Netz, D. J., Pierik, A. J., Stumpfig, M., Muhlenhoff, U. & Lill, R. The Cfd1-Nbp35 complex acts as a scaffold for iron-sulfur protein assembly in the yeast cytosol. *Nature Chem. Biol.* **3**, 278–286 (2007).
18. Abrahamsen, M. S. *et al.* Complete genome sequence of the apicomplexan, *Cryptosporidium parvum. Science* **304**, 441–445 (2004).
19. Loftus, B. *et al.* The genome of the protist parasite *Entamoeba histolytica. Nature* **433**, 865–868 (2005).
20. Carlton, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis. Science* **315**, 207–212 (2007).
21. Ali, V., Shigeta, Y., Tokumoto, U., Takahashi, Y. & Nozaki, T. An intestinal parasitic protist, *Entamoeba histolytica*, possesses a non-redundant nitrogen fixation-like system for iron-sulfur cluster assembly under anaerobic conditions. *J. Biol. Chem.* **279**, 16863–16874 (2004).
22. Johnson, D. C., Dean, D. R., Smith, A. D. & Johnson, M. K. Structure, function, and formation of biological iron-sulfur clusters. *Annu. Rev. Biochem.* **74**, 247–281 (2005).
23. Rouault, T. A. & Tong, W. H. Iron-sulphur cluster biogenesis and mitochondrial iron homeostasis. *Nature Rev. Mol. Cell Biol.* **6**, 345–351 (2005).
24. Balk, J. & Lobreaux, S. Biogenesis of iron-sulfur proteins in plants. *Trends Plant Sci.* **10**, 324–331 (2005).
25. Sutak, R. *et al.* Mitochondrial-type assembly of FeS centers in the hydrogenosomes of the amitochondriate eukaryote *Trichomonas vaginalis. Proc. Natl Acad. Sci. USA* **101**, 10368–10373 (2004).
26. Vavra, J. "Polar vesicles" of microsporidia are mitochondrial remnants ("mitosomes")? *Folia Parasitol. (Praha)* **52**, 193–195 (2005).
27. Muhlenhoff, U. *et al.* Functional characterization of the eukaryotic cysteine desulfurase Nfs1p from *Saccharomyces cerevisiae. J. Biol. Chem.* **279**, 36906–36915 (2004).

METHODS

Growth of *E. cuniculi* and *T. hominis*. Rabbit kidney cells (RK-13) infected with *E. cuniculi* strain II (provided by E. Didier) or *T. hominis* were grown as previously described[5].

Cloning of *T. hominis* NFS1, ISD11, ISU1 and YFH1. Degenerate PCR primers were used to amplify partial fragments of *T. hominis* *ThNFS1* and *ThISU1*, then the 5′ and 3′ ends and flanking sequences were completed using RACE (Marathon cDNA Amplification Kit, Clontech) and RAGE techniques[5]. The complete *ThNFS1* and *ThISU1* open reading frames were confirmed using *T. hominis* genomic DNA as a template for PCR. The genes for frataxin (*ThYFH1*) and Isd11 (*ThISD11*) were obtained from an ongoing project to sequence the *T. hominis* genome.

Yeast strains, cell growth and Fe–S protein detection methods. *Saccharomyces cerevisiae* strain W303-1A (*MAT**a**, ura3-1, ade2-1, trp1-1, his3-11, 15, leu2-3, 112*) was used as the reference wild-type strain; Gal-ISU1/Δisu2 (ref. 15), Gal-SSQ1 (ref. 15), Gal-GRX5 (ref. 15) and Gal-YFH1 (ref. 28) derivatives served for complementation studies. Yeast cell growth in rich or minimal media containing the appropriate carbon sources, radiolabelling experiments of Fe–S proteins using ^{55}Fe, and the measurement of mitochondrial and cytosolic enzyme activities were performed as described[12].

Plasmids for expression of microsporidian genes in yeast. The *T. hominis* genes *ThISU1*, *ThNFS1* and *ThmHSP70* were cloned into the yeast vector p426GPD, which contains the strong constitutive promoter of the glyceraldehyde-3-phosphate dehydrogenase gene[29]. Targeting of *T. hominis* proteins into yeast mitochondria was achieved by cloning the N-terminally truncated proteins behind the mitochondrial *Neurospora crassa* F_1-ATPase subunit β presequence[14]. The *E. cuniculi* Fe–S cluster genes were cloned into vector p426GPD containing or lacking the coding information for the mitochondrial presequence of subunit 9 of *N. crassa* F_o-ATPase[14]. Yeast *ISU1* was cloned in vector p424GPD.

Protein expression in *E. coli*. *ThISU1*, *ThNFS1* and *ThYFH1* were cloned into the pQE-40 expression vector (Qiagen) and expressed in *E. coli* M15 [pREP4] cells as recombinant histidine-tagged native proteins, or as dihydrofolate reductase (DHFR)-fusion proteins to enhance the efficiency of expression. In initial experiments *T. hominis* native ThNfs1 was found to be poorly immunogenic, so it was also expressed as a synthetic peptide (Nfs peptides) comprising the most variable domains of the protein. Expressed proteins were purified by gel electrophoresis for the (commercial) generation of mouse, rabbit or rat antisera.

ThNFS1 and *ThISD11* encoding an N-terminal His tag were cloned into pET-DUET1 (Novagen). After expression, a cell extract was produced and membranes were isolated by centrifugation at 150,000*g* for 30 min.

Western blotting. Total cell extracts were prepared from *T. hominis* purified spores, RK-13 cells infected with *T. hominis*, or uninfected RK-13 cells as a control. Extracts were blotted onto nitrocellulose membrane and blocked using standard protocols. Membranes were incubated with mouse anti-ThIsu1 (1:500), rabbit anti-ThmHsp70 (1:1,000), rabbit anti-ThNfs1 (1:1,000) or rabbit anti-ThYfh1 (1:500) followed by secondary anti-mouse or anti-rabbit antibodies conjugated to peroxidase (Molecular Probes) and detected using ECL.

Immunolocalization of frataxin, Isu1, Nfs1 and mitochondrial Hsp70 homologues in *E. cuniculi* and *T. hominis*. Rabbit kidney (RK-13) cells infected with either *E. cuniculi* or *T. hominis* were grown on coverslips until confluent and then fixed in acetone/methanol 50:50 v/v at −20 °C for 2 h before further processing for IFA. Meronts of *E. cuniculi* were also concentrated by disruption of infected RK-13 cells followed by filtration through glass wool[30]. The concentrated meronts were then placed onto poly-L-lysine-coated coverslips. After blocking with 5% milk in PBS, slides were incubated with a 1% milk-PBS solution with the relevant antiserum, and then subsequently incubated with goat fluorescent dye-conjugated secondary antibodies for 1 h. In assays in which two antibodies were raised in the same host (rabbit), we first purified (Melon IgG spin Purification kit; Pierce) and then biotinylated one of the antibodies (EZ-link Sulfo NHS-LC biotinylation kit; Pierce) to enable detection via a streptavidin conjugate. The fluorescent dyes used were Alexa 633 (red), Alexa 594 (red), or Alexa 488 (green) (Invitrogen). The labelled cells were visualized under a laser scanning confocal microscope (Leica TCS SP2 UV) using a ×63 lens (Plan Apo 1.32 nA). Images were collected using the manufacturer's software (Leica Microsystems). Software and images were processed and merged using Adobe Photoshop CS2.

28. Muhlenhoff, U., Richhardt, N., Ristow, M., Kispal, G. & Lill, R. The yeast frataxin homolog Yfh1p plays a specific role in the maturation of cellular Fe–S proteins. *Hum. Mol. Genet.* **11**, 2025–2036 (2002).

29. Mumberg, D., Muller, R. & Funk, M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* **156**, 119–122 (1995).

30. Taupin, V., Metenier, G., Vivares, C. P. & Prensier, G. An improved procedure for percoll gradient separation of sporogonial stages in *Encephalitozoon cuniculi* (Microsporidia). *Parasitol. Res.* **99**, 708–714 (2006).

LETTERS

Functional metagenomic profiling of nine biomes

Elizabeth A. Dinsdale[1,5]*, Robert A. Edwards[1,2,3,6]*, Dana Hall[1], Florent Angly[1,4], Mya Breitbart[7], Jennifer M. Brulc[8], Mike Furlan[1], Christelle Desnues[1]†, Matthew Haynes[1], Linlin Li[1], Lauren McDaniel[7], Mary Ann Moran[10], Karen E. Nelson[11], Christina Nilsson[12], Robert Olson[6], John Paul[7], Beltran Rodriguez Brito[1,4], Yijun Ruan[12], Brandon K. Swan[13], Rick Stevens[6], David L. Valentine[13], Rebecca Vega Thurber[1], Linda Wegley[1], Bryan A. White[8,9] & Forest Rohwer[1,2]

Microbial activities shape the biogeochemistry of the planet[1,2] and macroorganism health[3]. Determining the metabolic processes performed by microbes is important both for understanding and for manipulating ecosystems (for example, disruption of key processes that lead to disease, conservation of environmental services, and so on). Describing microbial function is hampered by the inability to culture most microbes and by high levels of genomic plasticity. Metagenomic approaches analyse microbial communities to determine the metabolic processes that are important for growth and survival in any given environment. Here we conduct a metagenomic comparison of almost 15 million sequences from 45 distinct microbiomes and, for the first time, 42 distinct viromes and show that there are strongly discriminatory metabolic profiles across environments. Most of the functional diversity was maintained in all of the communities, but the relative occurrence of metabolisms varied, and the differences between metagenomes predicted the biogeochemical conditions of each environment. The magnitude of the microbial metabolic capabilities encoded by the viromes was extensive, suggesting that they serve as a repository for storing and sharing genes among their microbial hosts and influence global evolutionary and metabolic processes.

Genomic plasticity of microbes causes variations in the gene content of closely related strains[4], making predictions of community metabolism on the basis of representative genomes and signature genes such as 16S ribosomal RNA unreliable. Although it seems that core genomes are relatively stable and shared among most individuals of the same species, parts of the genome (for example, prophages, CRISPRs, pathogenicity/ecological islands, ORFans) are hyper-variable[5]. Together, these two components make up the pan-genome[4]. Unlike the signature genes approach, metagenomic approaches analyse the complete genetic information of microbial and viral communities[6,7]. In this way, the relative abundances of all genes can be determined and used to generate a description of the functional potential of each community[8–14].

Here we use a comparative metagenomic approach to statistically analyse the frequency distribution of 14,585,213 microbial and viral metagenomic sequences to elucidate the functional potential of nine biomes including: subterranean (that is, mine samples); hypersaline ponds from solar salterns; marine; freshwater; coral-associated; microbialites (including stromatolites and thrombolites); aquaculture-fish-associated; terrestrial-animal-associated; and

mosquito-associated (details in Supplementary Table 1 and Supplementary Fig. 1). Microbial and viral metagenomes (Supplementary Fig. 2 and Supplementary Table 2) were isolated and pyrosequenced. The sequences were compared to the 2007 SEED platform (http://www.theseed.org) using the BLASTX algorithm, and hits with an E-value of <0.001 were considered to be significant (Methods). A total of 1,040,665 sequences from the 45 microbial metagenomes and 541,979 sequences from the 42 viral metagenomes were significantly similar to functional genes within the SEED (Supplementary Table 1). The SEED arranges metabolic pathways into a hierarchical structure in which all of the genes required for a specific task are arranged into subsystems[15]. At the highest level of organization, the subsystems include both catabolic and anabolic functions (for example, DNA metabolism) and at the lowest levels the subsystems are specific pathways (for example, the synthesis pathway for thymidine).

Table 1 shows the relative abundances of sequences assigned to each major subsystem in the combined analysis of the microbiomes

Table 1 | Mean percentage of sequences (± s.e.m.) similar to major metabolisms

Metabolic category	Microbial metagenomes	Viral metagenomes
Carbohydrates	17.218 (± 0.648)	14.353 (± 0.718)
Amino acids	12.036 (± 0.491)	10.132 (± 0.642)
Virulence	9.788 (± 0.339)	11.175 (± 0.508)
Protein metabolism	9.123 (± 0.497)	8.838 (± 0.522)
Respiration	7.139 (± 1.285)	3.718 (± 0.276)
Photosynthesis	6.965 (± 2.148)	1.984 (± 0.554)
Cofactors, vitamins, and so on	5.411 (± 0.226)	6.661 (± 0.393)
RNA metabolism	3.971 (± 0.195)	4.324 (± 0.387)
DNA metabolism	3.970 (± 0.180)	7.555 (± 0.943)
Nucleosides and nucleotides	3.316 (± 0.149)	7.666 (± 0.817)
Cell wall and capsule	3.235 (± 0.223)	5.098 (± 0.649)
Fatty acids and lipids	3.095 (± 0.160)	3.002 (± 0.242)
Membrane transport	2.736 (± 0.158)	2.371 (± 0.182)
Stress response	2.599 (± 0.115)	3.354 (± 0.326)
Aromatic compounds	2.351 (± 0.175)	2.550 (± 0.340)
Cell division and cell cycle	1.791 (± 0.091)	1.983 (± 0.212)
Nitrogen metabolism	1.547 (± 0.070)	1.135 (± 0.093)
Sulphur metabolism	1.230 (± 0.102)	1.302 (± 0.134)
Motility and chemotaxis	1.022 (± 0.096)	1.011 (± 0.083)
Phosphorus metabolism	0.909 (± 0.080)	1.319 (± 0.167)
Cell signalling	0.885 (± 0.076)	0.885 (± 0.072)
Potassium metabolism	0.796 (± 0.048)	0.846 (± 0.079)
Secondary metabolism	0.159 (± 0.014)	0.235 (± 0.047)

[1]Department of Biology, [2]Center for Microbial Sciences, [3]Department of Computer Sciences, and [4]Computational Science Research Centre, San Diego State University, San Diego, California 92182, USA. [5]School of Biological Sciences, Flinders University, Adelaide, South Australia 5042, Australia. [6]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. [7]University of South Florida, College of Marine Science, 140 7th Avenue South, St Petersburg, Florida 33701, USA. [8]Department of Animal Sciences, and [9]The Institute for Genomic Biology, University of Illinois, Urbana, Illinois 61801, USA. [10]Department of Marine Sciences, University of Georgia, Athens, 30602 Georgia, USA. [11]The J. Craig Venter Institute, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [12]Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [13]Department of Earth Science, University of California Santa Barbara, Santa Barbara, California 93106, USA. †Present address: Unité des Rickettsies, CNRS-UMR 6020, Faculté de médecine, 13385 Marseille, France.
*These authors contributed equally to this work.

Table 2 | Mean functional diversity and evenness (± s.e.m.) of metagenomes, sampled from nine environments

Biome	Functional diversity (H')		Functional evenness	
	Microbial	Viral	Microbial	Viral
Subterranean	2.393 (± 0.030)		0.005 (± 1.2×10^{-4})	
Hypersaline	2.361 (± 0.006)	2.041 (± 0.021)	0.005 (± 1.4×10^{-4})	0.012 (± 5.6×10^{-4})
Marine	2.313 (± 0.021)	2.162 (± 0.026)	0.005 (± 0.9×10^{-4})	0.007 (± 4.0×10^{-4})
Freshwater	2.430 (± 0.003)	2.080 (± 0.034)	0.005 (± 0.9×10^{-4})	0.010 (± 6.7×10^{-4})
Coral	1.733 (± 0.059)	2.289 (± 0.023)	0.009 (± 5.2×10^{-4})	0.007 (± 1.1×10^{-4})
Microbialites	2.408 (± 0.015)	1.743 (± 0.115)	0.005 (± 3.8×10^{-4})	0.019 (± 6.9×10^{-3})
Fish	2.447 (± 0.001)	2.439 (± 3.1×10^{-4})	0.005 (± 0.4×10^{-4})	0.005 (± 0.7×10^{-4})
Terrestrial animals	2.428 (± 0.006)	2.016 (± 0.173)	0.004 (± 0.1×10^{-4})	0.017 (± 4.5×10^{-3})
Mosquito		2.395 (± 0.015)		0.004 (± 0.5×10^{-4})

There are no subterranean viral metagenomes and no mosquito microbial metagenomes.

compared with the viromes. Over 30% of the identifiable genes in the microbiomes were associated with carbohydrate or protein metabolism. Respiration and photosynthesis subsystems accounted for an additional ~15% of the similarities. Subsystems responsible for nucleic acid metabolism and virulence were overrepresented in the viral fractions (Table 1), whereas respiration and photosynthesis genes were less frequent.

The functional diversity represented by the metagenomes approached its theoretical limit of 2.81 in all environments (Table 2), showing that most subsystems were represented in all of the samples. Only the coral-associated microbes showed a lower functional diversity; this is because they have fewer secondary metabolisms, virulence pathways, cell signalling pathways and membrane transport pathways. Because microbes associated with corals are taxonomically diverse[11], functional reduction may have occurred in these communities, similar to microbes in other symbiotic relationships[16].

Diversity is a function of both richness (that is, the number of metabolic processes) and evenness (that is, the relative abundance of a particular metabolic process in a sample). The evenness for the metagenomes was very low (<0.1; Table 2 and Supplementary Fig. 3), showing that there are a few dominant metabolisms in each environment. Differential dominant metabolisms suggest that there are characteristic functional profiles of the metagenomes.

To test the hypothesis that each environment has a distinguishing metabolic profile, a canonical discriminant analysis (CDA) was conducted (Fig. 1). Most of the variance between the different environments (79.8% of the combined microbiome and 69.9% of the virome) was explained in this analysis, showing that metagenomes are highly predictive of metabolic potential within an ecosystem. In contrast, a recent analysis of 16S rRNA genes from multiple environments only explained about 10% of the variance[17], suggesting that different ecosystems cannot be distinguished by their taxa.

The position of each metagenome in Fig. 1 reflects the frequency combination of sequences associated with each subsystem; the vectors indicate which metabolisms most strongly determined the distribution. Using these as clues, it is possible to determine which metabolisms are important for the organisms in that environment relative to other environments. For example, subsystems involved in respiration and protein metabolism placed the coral-associated microbes apart from the microbes found within terrestrial animals. This trend is visualized in Fig. 2, which shows that ~20% of the coral-associated microbial genes were involved in respiration, compared with only 3% in the microbiomes associated with terrestrial animals. The relatively high occurrence of respiration-associated genes in the coral-associated microbiomes reflects the diurnally fluctuating oxygen environment, which is supersaturated with oxygen in the day and essentially anaerobic at night[18]. In contrast, microbes living within the stable anaerobic alimentary tracts of terrestrial animals are less likely to experience selection for multiple respiration pathways.

Similarly, virulence genes were proportionally more abundant in the organism-associated microbes than in free-living microbes. These are the factors necessary to facilitate symbiotic relationships (mutualism, parasitism or commensalisms; Fig. 2f–h). Another

example of the predictive power of the metagenomes is the sulphur metabolisms associated with aquaculture fish. In particular, two subsystems—alkanesulphonate and taurine metabolism—were overrepresented in fish-associated metagenomes (Supplementary Fig. 4). Alkanesulphonates are involved in the use of both inorganic and organic sulphur, such as taurine and aliphatic sulphonates[19] (taurine is a sulphur organic acid used to supplement aquaculture fish food[20]).
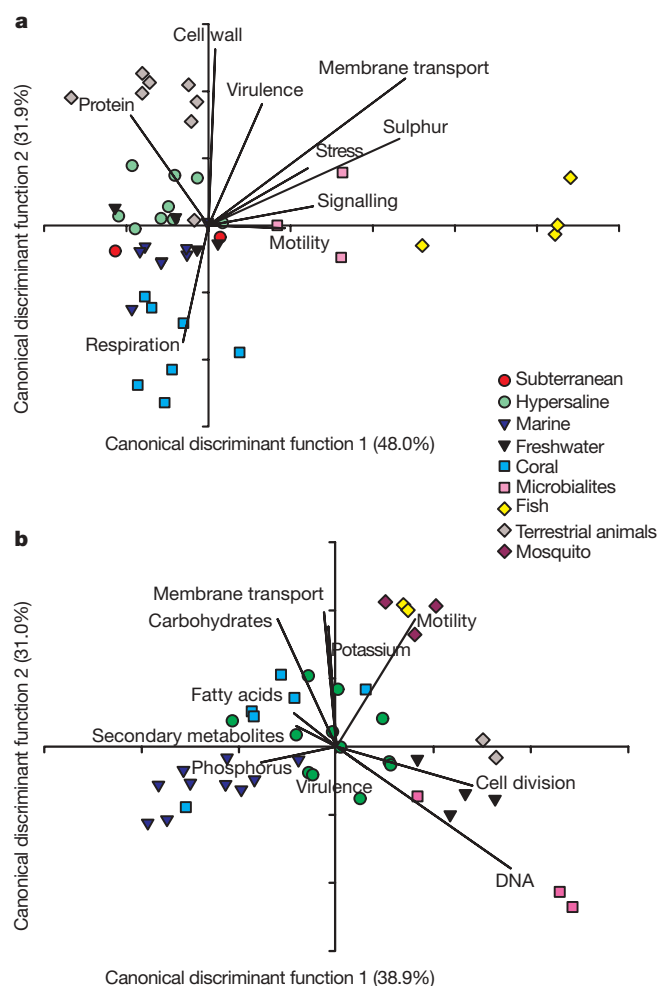
Figure 1 | Functional analysis of microbial and viral metagenomes. The CDA of the microbial (**a**) and viral (**b**) metagenomes identified that the metabolic processes grouped these communities in the two-dimensional spaced described by canonical discriminant functions 1 and 2. The symbols represent the position of each metagenome and the vectors represent the structural matrix for subsystems that were identified as influencing the separation of the metagenomes using the stepwise procedure. The length of the vectors represents the strength of influence of the particular metabolic process. The cross-validation scores for the microbial and viral metagenomes were 66.7 and 59.9%, respectively.

Together, these examples show that metagenomes predict important, emergent biological characters of the environments. By substituting environmental groups in multiple CDAs, the predictive nature of metagenomes was confirmed (Supplementary Figs 5 and 6).

Shifting of a metagenome from its sister group in the CDA was also predictive of ecological differences. For example, one of the marine metagenomes (number 27 Supplementary Table 1) was positioned more negatively than the rest of the marine metagenomes (Fig. 1a). This sample was taken from waters that were unusually rich in nitrogen, phosphate and dissolved organic carbon[21]. The ability to determine subtle differences in metabolic potential will allow the detection of environmental changes at early stages of perturbation and identify previously unknown pathways for therapeutics.

The viromes are dominated by phage, which are expected to have similar lifestyles in every environment (infection, replication, host lysis and release of free virions). Phage have also been shown to move between environments[22], which suggests that their metabolic profiles are similar in different ecosystems. In contrast, other studies have shown that phage carry 'specialization' genes[23], including phosphate metabolism[24] and cyanobacterial photosystems[25], to manipulate host metabolisms associated with a particular ecosystem. Phage 'sample' their host's genetic material and incorporate extra pieces of DNA called MORONS[26], suggesting that phage metagenomes may instead show distinctive profiles based on their environment. As shown in Figs 1b and 2, the viromes have highly predictive metabolic profiles that suggest enrichment for specific genes in different environments, and thus support the latter hypothesis (69.9% of the variance).

Because phages and viruses are non-motile, the abundance of motility and chemotaxis proteins within the combined viral metagenomes was the most unexpected example of specialized metabolisms being carried within the viromes (Fig. 3). A total of 130 SEED-annotated motility and chemotaxis proteins (out of a possible 157) were present in the viromes. There was a non-random acquisition of these proteins by the viral community, shown by the variation in relative abundances of these proteins between the microbial and viral metagenomes (Supplementary Table 3). In the viromes, flagellar biosynthesis protein FlhA, the chemotaxis response regulator proteins CheA and CheB and deacylases were overrepresented (Supplementary Table 3), whereas the twitching motility protein PilT, type II secretary pathways and GldJ were overrepresented in the microbiomes. *cheA* and *cheB* genes within microbes work together to control flagella motor switching rates[27], but their role within the phage remains an outstanding question.

Essentially all of the functional diversity was represented in the viromes. Unlike their cellular hosts, most viruses must carry a specific amount of DNA to correctly pack their capsids (that is, viruses are not evolutionarily penalized for carrying 'extra' DNA). If there is a selective advantage of the extra DNA (resulting in increased phage progeny), these genes are fixed in the phage genome; otherwise they will be lost. Because there are an estimated 10^{31} phages on the planet and they can move between environments, the potential reservoir of genes that can be transferred both locally and globally[12] by phage is enormous[28]. As our research shows, there is little restriction to the types of genes carried by the viral community, suggesting that they influence a wide range of processes, including biogeochemical cycling, short-term adaptation and long-term evolution of microbes.

The low functional evenness measured for both microbial and viral metagenomes is even lower than the functional diversity calculated for individual bacterial genomes (Table 2 and Supplementary Fig. 3). This finding is diametrically opposed to the high taxonomic evenness reported for both microbial and viral communities[2,12], ranging from 0.6 to 1 for human faecal and marine viruses[9,12] and about 0.9 for soil microbes[29]. To resolve this apparent dilemma, we propose that the frequency of a gene encoding a particular metabolic function reflects its relative importance in an environment, and that genetic sweeps favour particular gene frequencies regardless of their taxonomical background. That is, rather than changing taxa, variation in gene content, presumably by means of horizontal gene transfer[30] between
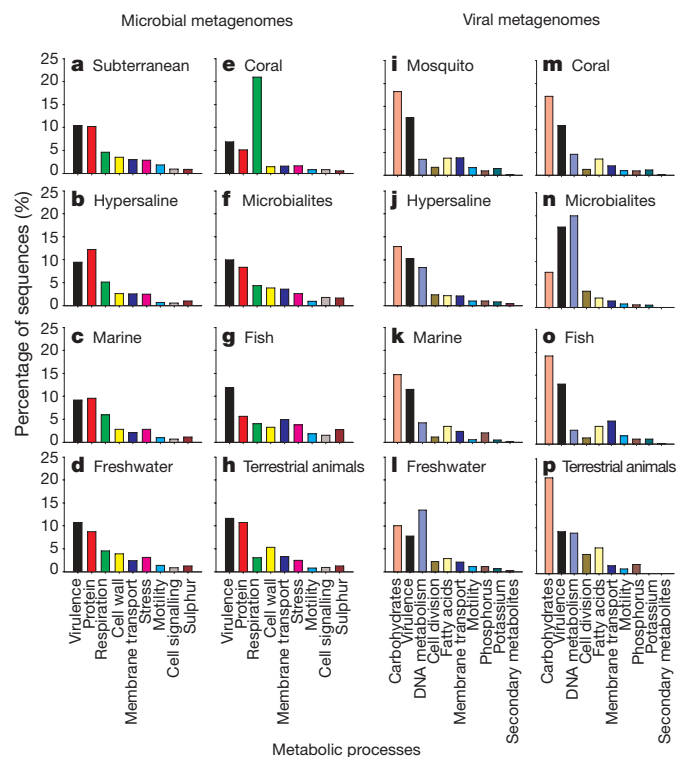
Figure 2 | A one-dimensional representation of the environmental metabolic profiles for the microbial and viral metagenomes sampled from the nine environments. Microbial metagenomes are shown in **a–h**, and viral metagenomes are shown in **i–p**. Each bar represents the mean for each metabolic category. For subterranean, n = 2 (**a**); for hypersaline, n = 9 (**b**); for marine, n = 8 (**c**); for freshwater, n = 4 (**d**); for coral, n = 7 (**e**); for microbialites, n = 3 (**f**); for fish, n = 4 (**g**); for terrestrial animals, n = 8 (**h**); for mosquito, n = 3 (**i**); for hypersaline, n = 12 (**j**); for marine, n = 10 (**k**); for freshwater n = 4 (**l**); for coral n = 6 (**m**); for microbialites, n = 3 (**n**); for fish, n = 2 (**o**); and for terrestrial animals, n = 2 (**p**).
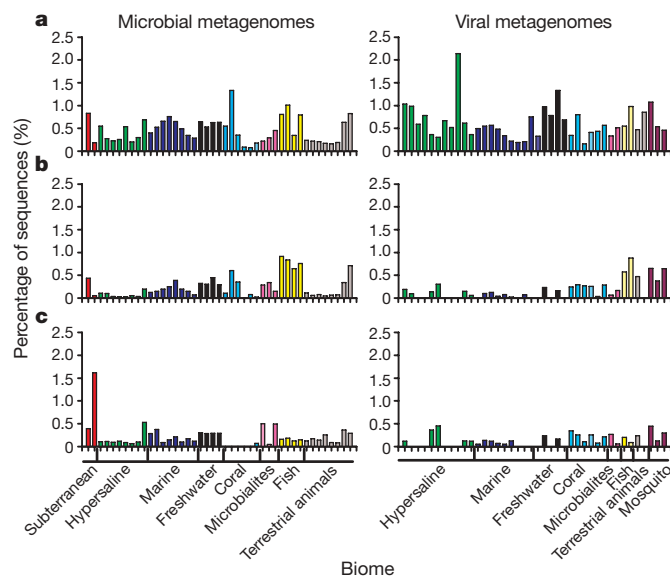
Figure 3 | A comparison of the distribution of sequences similar to motility and chemotaxis genes identified within the microbiomes (n = 43) and viromes (n = 41). Microbial metagenomes are shown on the left, and viral metagenomes are shown on the right. The abundance of sequences identified within each of three fine-scale subsystems including flagellum (**a**), bacterial chemotaxis (**b**) and gliding motility (**c**), as described by the SEED platform.

sympatric microbes, is controlling gene distribution within an environment. The large amount of variation (~70%) explained by the functional analysis presented here supports this hypothesis.

METHODS SUMMARY

Samples for metagenomes were collected and fractioned using standard techniques, sequenced using pyrosequencing and compared to the functional genes in the SEED platform[11,12] (Methods). All statistics were performed on the percentage of sequences showing similarities to known functions. For the CDA, sequences were grouped according to the SEED classification scheme and the analysis was conducted on the principal metabolic functions. The CDA builds a model for group membership. A discriminative value is calculated for each metagenomic sample, which is a linear combination of the response variables (metabolic processes) represented in the new dimensional space. These values are used to visualize group membership.

An advantage of the CDA is that it identifies which variables are driving the separation between the groups; it uses these to build the model and discards those that are not influential. Identification of influential variables was conducted by a stepwise method, using Wilk's lambda with $P = 0.05$, and was confirmed with analysis of variance (ANOVA; Supplementary Table 4). The level of influence of each variable is provided by the structural matrix and can be visualized using an h-plot, in which the length of the line is representative of the level of influence. The CDA also performs a cross-validation analysis that identifies the likelihood of correctly classifying each sample. Cross validation removes the predetermined grouping for each sample and uses the response variables to align the individual sample to a group. Because the data were divided into nine predetermined groups (biomes), the number of samples correctly identified by chance alone is 11%. The percentage-correct classification has to be substantially larger than this number for the metabolic processes to be useful for classifying the metagenomes into environments.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Newman, D. K. & Banfield, J. F. Geomicrobiology: how molecular-scale interactions underpin biogeochemical systems. *Science* **296**, 1071–1076 (2002).
2. Prosser, J. I. *et al.* The role of ecological theory in microbial ecology. *Nature Rev. Microbiol.* **5**, 384–392 (2007).
3. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
4. Medini, D. *et al.* The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
5. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of *Prochlorococcus. Science* **311**, 1768–1770 (2006).
6. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006).
7. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nature Rev. Genet.* **6**, 805–814 (2005).
8. Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57 (2006).
9. Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).
10. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99**, 14250–14255 (2002).
11. Wegley, L., Breitbart, M., Edwards, R. A. & Rohwer, F. Metagenomic analysis of the microbial community associated with the coral *Porites astreoides. Environ. Microbiol.* **9**, 2707–2719 (2007).
12. Angly, F. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).
13. Breitbart, M. & Rohwer, F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* **39**, 729–736 (2005).
14. Fierer, N. *et al.* Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of Bacteria, Archaea, Fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73**, 7059–7066 (2007).
15. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
16. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Microbiol.* **17**, 589–596 (2001).
17. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl Acad. Sci. USA* **104**, 11436–11440 (2007).
18. Shashar, N., Cohen, Y. & Loya, Y. Extreme diel fluctuations of oxygen in diffusive boundary layers surrounding stony corals. *Biol. Bull.* **185**, 455–461 (1993).
19. Iwanicka-Nowicka, R. *et al.* Regulation of sulfur assimilation pathways in *Burkholderia cenocepacia*: identification of transcription factors CysB and SsuR and their role in control of target genes. *J. Bacteriol.* **189**, 1675–1688 (2007).
20. Aksnes, A., Hope, B., Hostmark, O. & Albrektsen, S. Inclusion of size fractionated fish hydrolysate in high plant protein diets for Atlantic cod, *Gadus morhua. Aquaculture* **261**, 1102–1110 (2006).
21. Dinsdale, E. A. *et al.* Microbial ecology of four coral atolls in the Northern Line Islands. *Plos One* **3**, e1584 (2008).
22. Sano, E., Carlson, S., Wegley, L. & Rohwer, F. Movement of virus between biomes. *Appl. Environ. Microbiol.* **70**, 5842–5846 (2004).
23. Davis, B. M. & Waldor, K. in *Mobile DNA II* (ed. Craig, N. L., Gragie, R., Gellert, M. & Lambowitz, A. M.) 1040–1055 (ASM, Washington DC, 2002).
24. Rohwer, F. *et al.* The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45**, 408–418 (2000).
25. Mann, N. *et al.* Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424**, 741 (2003).
26. Hendrix, R. W., Smith, M. C. M., Burns, R. N. & Ford, M. E. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA* **96**, 2192–2197 (1999).
27. Wadhams, G. H. & Armitage, J. P. Making sense of it all: bacterial chemotaxis. *Nature Rev. Mol. Cell Biol.* **5**, 1024–1037 (2004).
28. Hendrix, R. W. Bacteriophage: evolution of the majority. *Theor. Popul. Biol.* **61**, 471–480 (2002).
29. Dunbar, J., Ticknor, L. O. & Kuske, C. R. Assessment of microbial diversity in four southwestern United States soils by 16S rRNA gene terminal restriction fragment analysis. *Appl. Environ. Microbiol.* **66**, 2943–2950 (2000).
30. Frigaard, N.-U., Martinez, A., Mincer, T. J. & Delong, E. F. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**, 847–850 (2006).

Author Contributions E.A.D. conceptualized the project, conducted the CDA and wrote the manuscript. R.A.E., R.O. and R.S. performed the bioinformatics. D.H. conducted the non-parametric statistical analysis. F.R. oversaw most of the metagenomic projects. All other authors collected the metagenomes and provided comments on the manuscript.

Author Information The metagenomes used in this paper are freely available from the SEED platform and are being made accessible from CAMERA and the NCBI Short Read Archive. The accession numbers are shown in Supplementary Table 1. The NCBI genome project IDs used in this study are: 28619, 28613, 28611, 28609, 28607, 28605, 28603, 28601, 28599, 28597, 28469, 28467, 28465, 28463, 28461, 28459, 28457, 28455, 28453, 28451, 28449, 28447, 28445, 28443, 28441, 28439, 28437, 28435, 28433, 28431, 28429, 28427, 28425, 28423, 28421, 28419, 28417, 28415, 28413, 28411, 28409, 28407, 28405, 28403, 28401, 28395, 28393, 28391, 28389, 28387, 28385, 28383, 28381, 28379, 28377, 28375, 28373, 28371, 28361, 28359, 28357, 28355, 28353 and 28351. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.A.D. (elizabeth_dinsdale@hotmail.com).

METHODS

Collection of the metagenomes. Metagenomic samples were collected and DNA was prepared by the different groups involved; each laboratory used slight modifications on the basic protocol. Sample locations were widely dispersed or separate organisms (Supplementary Fig. 1 and Supplementary Table 1). Metagenomes were collected to answer broad ecological questions such as viral community dynamics in the lungs of healthy and cystic fibrosis patients and the microbial communities on coral reefs (Supplementary Table 1). Typically, the microbiome process starts by filtering samples onto 0.22 μm Sterivex filters, removing the filter membranes and extracting DNA using a bead-beating protocol (MoBio). In some samples, the DNA was amplified with Genomiphi (GE Healthcare Life Sciences) in six to eight 18-h reactions[22,31]. The reactions were pooled and purified using silica columns (Qiagen). The DNA was precipitated with ethanol and resuspended in water at a concentration of approximately 300 ng μl^{-1}. Microbial metagenomes capture Bacteria, Archaea, some small protists as well as a few trapped viral-like particles (Supplementary Table 2).

The viruses in the small metagenomic fractions (that is, 0.22-μm filtrate treated with chloroform) were purified using caesium chloride (CsCl) step gradients to remove free DNA and any cellular material[10,12]. Viral samples were visually checked for microbial contamination using epifluorescent microscopy. Viral DNA was isolated using CTAB (cyltrimethylammonium bromide) and 25:24:1 phenol:chloroform:isoamyl alcohol mix extractions and amplified using Genomiphi reactions. These reactions were pooled and purified using silica columns (Qiagen). The DNA was precipitated with ethanol and resuspended in water at a concentration of approximately 300 ng μl^{-1}. One viral metagenome (number 40, Supplementary Table 1) was prepared by concentrating a natural microbial sample and inducing it with mitomycin C. All metagenome libraries consisted of approximately 5 μg of DNA. The viral metagenomes contained viruses, phage and prophage, and as expected the proportion of phage and prophage are higher in these metagenomes than in the microbial fraction (Supplementary Table 2).

Sequencing and bioinformatics. Sequencing was performed using pyrosequencing on Roche Applied Sciences and 454 Life Sciences GS20 platforms[32] with a practical limit of 105 bp. DNA sequences were analysed in the metagenomics RAST pipeline—an open-access metagenome curation and analysis platform (http://metagenomics.theseed.org/)[33]. First, sequences were screened to remove exactly duplicated sequences—a known artefact of the pyrosequencing approach. The sequences were compared to the SEED platform, which comprises all known protein sequences, using the NCBI BLASTX algorithm on the NMPDR compute cluster (Argonne National Laboratory; http://www.nmpdr.org/). The SEED platform includes all available genome data, DNA and protein sequences, and is supplemented with data from genome sequencing centres as available. Every metagenome was compared to exactly the same data set using the same BLAST parameters at the same time to ensure congruity of the data. Connections between the metagenomes and the SEED subsystems were calculated by identifying matches to the SEED platform where the matched protein was curated to be in a subsystem, and the expect value from the BLAST search was less than 0.001. The SEED subsystems are manually curated collections of proteins with related functions and are available at http://www.theseed.org/. Simultaneously, all sequences were compared to the 16S databases using BLASTN. The databases were extracted from GreenGenes[34], the Ribosomal Database Project[35] and the European Ribosomal Database Project[36].

Several metagenomes were constructed from environments that were likely to contain DNA from other organisms such as humans, corals and mosquitoes. To test and to remove contaminants, 20,000 sequences were chosen at random from every metagenome and compared to the March 2006 build of the human genome and the February 2003 build of the *Anopheles gambiae* genome (both downloaded from http://genome.ucsc.edu/). The comparisons were performed using BLASTN with an expect (*E*) value cutoff of 1×10^{-5}. Every sample (including the mosquito samples) had less than 1% of their sequences with significant similarity to the *A. gambiae* genome, and only two samples had >5% of sequence similarity to the human genome. These two samples, from the human virome studies, were compared in full and human sequences excluded. To identify and remove dinoflagellate sequences, such as *Symbiodinium* (a coral symbiont), a custom database was created from the nucleotide and RNA (expressed sequence tag) sequences in GenBank; all coral reef water and coral samples were analysed as described above and dinoflagellates sequences were excluded.

Statistical analysis. Statistics were performed on the proportions of sequences within each subsystem, thus normalizing data across metagenomes and removing differences in reaction efficiencies. Total numbers of sequences and numbers of sequences that showed similarities to the SEED are provided in Supplementary Table 1, and ~11% of sequences were similar to functional genes. The SEED platform housed 654 well-documented subsystems that were used to calculate the Shannon index (H'). Maximum diversity occurs when every functional category is present in equal numbers, thus $H_{max} = \log S$, where S is number of categories. Evenness is H' divided by the number of subsystems in each sample (evenness ranges from 0 to 1, which is even). As a comparison to the metagenomic analyses, the diversity and evenness was calculated for all 842 sequenced bacterial genomes. These calculations were conducted on the number of genes within each subsystem, rather than on the number of sequences that was used for the metagenomes (Supplementary Fig. 3).

To analyse the stability of the CDAs, an experiment was conducted in which several of the metagenomic groups were removed and the analysis re-run. In the first trial, the subterranean, fish and mosquito metagenomes were removed (Supplementary Fig. 5). In the second trial, these metagenomes were re-added and the hypersaline metagenomes removed (Supplementary Fig. 6). Multiple trials were required because CDAs are sensitive to the number of samples (that is, metagenomes) relative to the number of variables (that is, metabolic processes).

The data were further analysed using a non-parametric ANOVA, a Kruskal–Wallis test and a median test, and the results compared to ensure that stable results could be obtained (Supplementary Table 3). Environments driving the variation were identified using Duncan comparisons (degrees of freedom were set at 7).

All metagenomes were provided by authors of this manuscript. Further material, including direct access to the data, is available at http://www.theseed.org/DinsdaleSupplementalMaterial/. The NCBI genome project IDs used in this study that were associated with previous publications are: 28369, 28367, 28365, 28363, 28349, 28347, 28345, 28343, 19145 17771, 17769, 17767, 17765 17635, 17633 and 17401.

31. Gunn, M. R. *et al.* A test of the efficacy of whole-genome amplification on DNA obtained from low-yield samples. *Mol. Ecol. Notes* **7**, 393–399 (2007).
32. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
33. Meyer, F. *et al.* The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinf.* (submitted).
34. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
35. Cole, J. R. *et al.* The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.* **35**, D169–D172 (2007).
36. Wuyts, J., Perriere, G. & de Peer, Y. V. The European ribosomal RNA database. *Nucleic Acids Res.* **32**, D101–D103 (2004).

CORRIGENDUM

Functional metagenomic profiling of nine biomes

Elizabeth A. Dinsdale, Robert A. Edwards, Dana Hall, Florent Angly, Mya Breitbart, Jennifer M. Brulc, Mike Furlan, Christelle Desnues, Matthew Haynes, Linlin Li, Lauren McDaniel, Mary Ann Moran, Karen E. Nelson, Christina Nilsson, Robert Olson, John Paul, Beltran Rodriguez Brito, Yijun Ruan, Brandon K. Swan, Rick Stevens, David L. Valentine, Rebecca Vega Thurber, Linda Wegley, Bryan A. White & Forest Rohwer

Nature **452**, 629–632 (2008)

In this Letter, functional diversity and evenness were calculated using $\log_{10}$, but it has been drawn to our attention that the calculations should have been made with natural logs (Table 2). Recalculation does not change the relative levels of diversity; however, this reanalysis showed that a mistake was made in the original evenness calculation. Corrected values are shown in Table 1 below.

Table 1 | Functional metagenomic diversity calculated using the natural log (*H*′ max 6.483) and corrected evenness (± s.e.m.)

Biome	Diversity natural log		Evenness	
	Microbial	Viral	Microbial	Viral
Subterranean	5.511 (± 0.098)		0.881 (± 0.008)	
Hypersaline	5.435 (± 0.043)	4.699 (± 0.169)	0.892 (± 0.006)	0.893 (± 0.013)
Marine	5.326 (± 0.134)	4.979 (± 0.188)	0.875 (± 0.021)	0.856 (± 0.015)
Freshwater	5.595 (± 0.015)	4.790 (± 0.155)	0.899 (± 0.004)	0.888 (± 0.015)
Coral	3.991 (± 0.363)	5.271 (± 0.130)	0.748 (± 0.040)	0.901 (± 0.015)
Microbialites	5.544 (± 0.059)	4.014 (± 0.460)	0.903 (± 0.011)	0.830 (± 0.085)
Fish	5.634 (± 0.007)	5.615 (± 0.001)	0.908 (± 0.002)	0.913 (± 0.003)
Terrestrial animals	5.590 (± 0.040)	4.642 (± 0.400)	0.886 (± 0.006)	0.965 (± 0.009)
Mosquito		5.514 (± 0.060)		0.875 (± 0.011)

A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25

Rayjean J. Hung[1,2]*, James D. McKay[1]*, Valerie Gaborieau[1], Paolo Boffetta[1], Mia Hashibe[1], David Zaridze[3], Anush Mukeria[3], Neonilia Szeszenia-Dabrowska[4], Jolanta Lissowska[5], Peter Rudnai[6], Eleonora Fabianova[7], Dana Mates[8], Vladimir Bencko[9], Lenka Foretova[10], Vladimir Janout[11], Chu Chen[12], Gary Goodman[12], John K. Field[13], Triantafillos Liloglou[13], George Xinarianos[13], Adrian Cassidy[13], John McLaughlin[14], Geoffrey Liu[15], Steven Narod[16], Hans E. Krokan[17], Frank Skorpen[17], Maiken Bratt Elvestad[17], Kristian Hveem[17], Lars Vatten[17], Jakob Linseisen[18], Françoise Clavel-Chapelon[19], Paolo Vineis[20,21], H. Bas Bueno-de-Mesquita[22], Eiliv Lund[23], Carmen Martinez[24], Sheila Bingham[25], Torgny Rasmuson[26], Pierre Hainaut[1], Elio Riboli[20], Wolfgang Ahrens[27], Simone Benhamou[28,29], Pagona Lagiou[30], Dimitrios Trichopoulos[30], Ivana Holcátová[31], Franco Merletti[32], Kristina Kjaerheim[33], Antonio Agudo[34], Gary Macfarlane[35], Renato Talamini[36], Lorenzo Simonato[37], Ray Lowry[38], David I. Conway[39], Ariana Znaor[40], Claire Healy[41], Diana Zelenika[42], Anne Boland[42], Marc Delepine[42], Mario Foglio[42], Doris Lechner[42], Fumihiko Matsuda[42], Helene Blanche[43], Ivo Gut[42], Simon Heath[43], Mark Lathrop[42,43] & Paul Brennan[1]

Lung cancer is the most common cause of cancer death worldwide, with over one million cases annually[1]. To identify genetic factors that modify disease risk, we conducted a genome-wide association study by analysing 317,139 single-nucleotide polymorphisms in 1,989 lung cancer cases and 2,625 controls from six central European countries. We identified a locus in chromosome region 15q25 that was strongly associated with lung cancer ($P = 9 \times 10^{-10}$). This locus was replicated in five separate lung cancer studies comprising an additional 2,513 lung cancer cases and 4,752 controls ($P = 5 \times 10^{-20}$ overall), and it was found to account for 14% (attributable risk) of lung cancer cases. Statistically similar risks were observed irrespective of smoking status or propensity to smoke tobacco. The association region contains several genes, including three that encode nicotinic acetylcholine receptor subunits (*CHRNA5*, *CHRNA3* and *CHRNB4*). Such subunits are expressed in neurons and other tissues, in particular alveolar epithelial cells, pulmonary neuroendocrine cells and lung cancer cell lines[2,3], and they bind to N′-nitrosonornicotine and potential lung carcinogens[4]. A non-synonymous variant of *CHRNA5* that induces an amino acid substitution (D398N) at a highly conserved site in the second intracellular loop of the protein is among the markers with the strongest disease associations. Our results provide compelling evidence of a locus at 15q25 predisposing to lung cancer, and reinforce interest in nicotinic acetylcholine receptors as potential disease candidates and chemopreventative targets[5].

Lung cancer is caused predominantly by tobacco smoking, with cessation of tobacco consumption being the primary method for prevention. The risk among those who quit smoking remains elevated (although less than those who continue to smoke), and former smokers make up an increasing proportion of lung cancer patients in countries where tobacco consumption has declined[6,7]. Treatment strategies are of limited efficacy, with an overall 5-year survival rate of about 15%[8]. Lung cancer has an important heritable component[9], and identifying genes that are involved may suggest chemoprevention targets or allow for identification of groups at high risk. Despite a large number of studies including both sporadic and multi-case families, success in identifying genes that cause lung cancer has been extremely limited.

The availability of tagging single-nucleotide polymorphism (SNP) panels across the whole genome allows for efficient and comprehensive analysis of common genomic variation to be conducted without a priori hypotheses based on gene function or disease pathways. They

633

require very large series of cases and controls to ensure adequate statistical power, and multiple subsequent studies to confirm the initial findings. We conducted a genome-wide association study of lung cancer using the Illumina Sentrix HumanHap300 BeadChip containing 317,139 SNPs and estimated to tag approximately 80% of common genomic variation[10]. We initially genotyped 1,989 cases and 2,625 controls from the International Agency for Research on Cancer (IARC) central Europe lung cancer study. This was conducted in six countries between 1998 and 2002 and each centre followed an identical protocol to recruit newly diagnosed cases of primary lung cancer, as well as a comparable group of population or hospital controls (Supplementary Methods). We excluded samples that failed one of several quality control criteria (Supplementary Methods) or because they showed evidence of admixture with Asian ethnicity (Supplementary Fig. 1); we also excluded 7,116 problematic SNPs. This resulted in a comparison of 310,023 SNPs between 1,926 cases and 2,522 controls.

We analysed each SNP individually by calculating P-values for trend in a logistic regression model and incorporating additional parameters including country, age and sex (Supplementary Methods). The distribution of the bottom 90% of P-values was similar to the expected distribution, and the genomic control parameter was 1.03, implying that there was no systematic increase in false-positive findings owing to population stratification or any other form of bias (Fig. 1a). However, there was a marked deviation between the observed and expected P-values among the top 10% (Fig. 1b). In particular, two SNPs on chromosome 15q25, rs1051730 and rs8034191, were strongly associated with disease ($P = 5 \times 10^{-9}$ and $P = 9 \times 10^{-10}$, respectively), exceeding the genome-wide significance level of $P = 5 \times 10^{-7}$ (Fig. 1c). Further analysis incorporating adjustment by principal components indicated that population stratification was unlikely to account for this observation (Supplementary Methods).

The odds ratio (OR) and 95% confidence interval (CI) for carrying one copy of the most significant marker (rs8034191), adjusted by age, sex and country, was 1.27 (1.11–1.44) and for carrying two copies of the allele was 1.80 (1.49–2.18); the allelic OR was 1.32 (1.21–1.45). When the data were analysed separately by country of origin, we found a significant association in all countries except Romania, which had the smallest sample numbers, although the trend in Romania was similar and the association was significant under a

dominant model (data not shown). There was no evidence of heterogeneity by country of origin ($P = 0.58$). Further adjustment was undertaken for various tobacco-related variables including duration of smoking, pack years (average number of cigarette packs per day multiplied by years of smoking) and age at onset of smoking. Adjustment by duration of smoking provided the best-fitting model to account for tobacco use based on the Akaike's information criteria (Supplementary Methods), although the adjusted estimates with duration of smoking (allelic OR = 1.28 (1.16–1.42)) were similar to the estimates adjusted by age, sex and country only.

We investigated further the association by genotyping 34 additional 15q25 markers that were selected as follows. First, we used an imputation method (see http://www.sph.umich.edu/csg/abecasis/MACH/index.html) to identify additional genetic variants from the Centre d'Etude du Polymorphism Humain Utah (CEU) HapMap data that are likely to have a strong disease association, but are not present in the HumanHap300 panel. We attempted genotyping of SNPs from the 15q25 region with an association P-value of the imputed data of $<10^{-6}$. Second, we included SNPs of CHRNA5 and CHRNA3 that had been included in a previous study of these genes in nicotine dependence[11]. Third, we attempted genotyping of all non-synonymous SNPs in dbSNP from the six genes within or near the association region. The results for all markers tested in the 15q25 region, including those in the HumanHap300 panel, are shown in Supplementary Table 1. Twenty-three of the additional genotyped markers showed evidence of association exceeding the genome-wide significance level of 5×10^{-7} (Fig. 2). These span more than 182 kilobases (kb) but are in strong linkage disequilibrium (pairwise $D' > 0.8$ and $r^2 > 0.6$) with two predominant haplotypes accounting for more than 85% of the haplotypes in patients and controls (Supplementary Table 2).

To confirm our findings we genotyped rs8034191 and rs16969968 (where rs16969968 is a second variant with a strong disease association) in five further independent studies of lung cancer: the European Prospective Investigation in Cancer and Nutrition (EPIC) cohort study (781 cases and 1,578 controls), the Beta-Carotene and Retinol Efficacy Trial (CARET) cohort study (764 cases and 1,515 controls), the Health Study of Nord-Trøndelag (HUNT) and Tromsø cohort studies (235 cases and 392 controls), the Liverpool lung cancer case-control study (403 cases and 814 controls), and the Toronto lung cancer case-control study (330 cases and 453 controls) (Supplementary Methods). We observed an increased risk for both heterozygous and homozygous variants of rs8034191 in all five replication samples (Table 1), with no evidence of any statistical heterogeneity between studies. After pooling across all six studies, the ORs (95% CI) were 1.21 (1.11–1.31) and 1.77 (1.58–2.00) for heterozygous and homozygous carriers, respectively, the allelic OR was 1.30 (1.23–1.37), and the P-value for trend was 5×10^{-20}. Further adjustment for duration of tobacco smoking did not alter the estimates: allelic OR = 1.30 (1.22–1.40). The genotype-specific model that estimated the OR for heterozygous and homozygous carriers separately was a significantly better fit than the model estimating the allelic OR ($P = 0.025$), suggesting a potential recessive effect.

The prevalence of the variant allele was 34%, resulting in 66% of the control participants carrying at least one copy, and the percentage of lung cancer explained by carrying at least one allele (that is, the population attributable risk) was 15% in the combined data set. We obtained a similar attributable risk in the central European study (16%) and in the replication studies (14%). The second variant with strong disease association (rs16969968) that was genotyped in the five replication studies gave very similar results, as expected from the strong linkage disequilibrium ($D' = 1.00$, $r^2 = 0.92$) among the disease-associated markers (allelic OR = 1.30 (1.23–1.38); $P = 1 \times 10^{-20}$).

The large number of patients in the combined data set allowed us to examine the association in different smoking categories and with respect to different histological subtypes (Supplementary Table 3 and
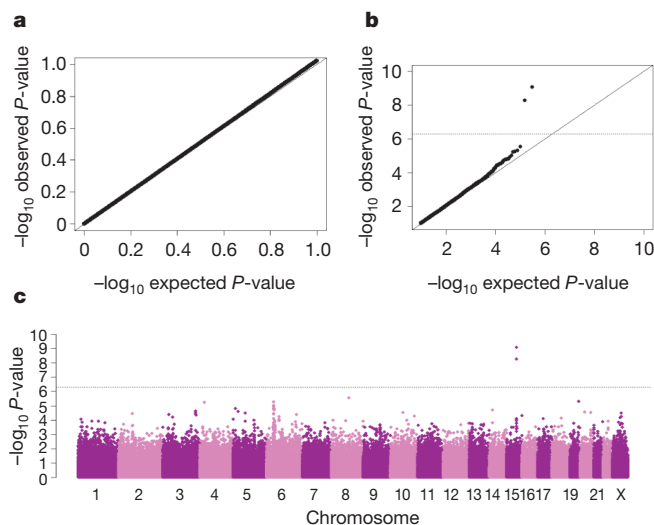

Figure 1 | Genome-wide association results in the central Europe study.
a–c, Quantile–quantile plot for bottom 90% of P-values (a) and top 10% of P-values (b), as well as scatter plot (c) of P-values in −log scale from the trend test for 310,023 genotyped variants comparing 1,926 lung cancer cases and 2,522 controls.
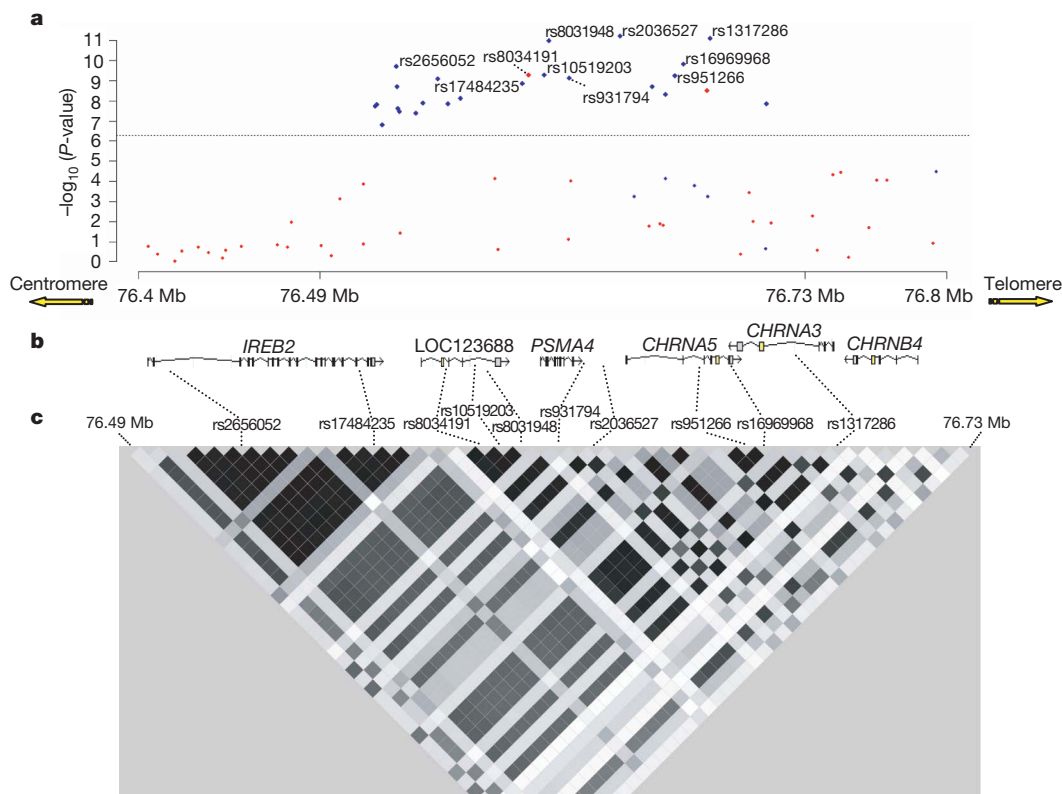
Figure 2 | Lung cancer area of interest across 15q25. a, *P*-values for SNPs genotyped in the 15q25 region (76.4–76.8 Mb). The dotted line indicates the genome-wide threshold of $P < 5 \times 10^{-7}$. Points labelled with rs numbers have a $P < 1 \times 10^{-9}$. Points in red are genotyped in the 317K Illumina panel; points in blue indicate additional genotyped SNPs (Taqman). **b**, Positions of the six known genes. **c**, Pairwise r^2 estimates for 46 common SNPs from 76.49 Mb to 76.73 Mb in controls from the central Europe IARC study, with increasing shades of grey indicating higher r^2 values. The majority of pairwise *D*′ estimates for these SNPs exceed 0.8.

Supplementary Discussion). Increased risks were seen for former smokers ($P = 4 \times 10^{-7}$) and current smokers ($P = 3 \times 10^{-10}$), as well as a potential increased risk for people who had never smoked ($P = 0.013$). No appreciable variation of the risk was found across the main histological subtypes of lung cancer. We observed a similar risk after stratifying by age at diagnosis, and a slightly greater risk for women compared to men ($P = 0.06$) (Supplementary Table 3). Analysis of the susceptibility locus in additional lung cancer studies would be desirable to obtain further information on these patterns of risk, particularly with respect to smoking status, cumulative cigarette consumption, age and sex. Notably, the risk haplotype is rare in Asian (Japanese and Chinese) and not observed in African (Yoruba) data in the HapMap database[12] and many of the risk alleles have markedly varied allele frequencies in different populations (Supplementary Table 1). Thus, future examination of the association of these markers with lung cancer in different populations might contribute to refined mapping of the locus.

We further investigated whether the locus was associated with cancers of the head and neck including those of the oral cavity, larynx, pharynx and oesophagus. We analysed rs8034191 in two separate studies of head and neck cancer conducted in Europe, the first being conducted in five countries of central Europe and overlapping with the lung cancer controls from five of the six countries included in the present genome-wide association study (726 cases and 694 controls), and the second study being conducted in eight countries of Europe (the ARCAGE study) and including 1,536 cases and 1,443 controls. We observed no effect in either of the two studies separately or combined or in any of the cancer subgroups (Supplementary Fig. 2), implying that this association was specific for lung cancer. Similar results were also observed for rs16969968 (data not shown).

The disease-associated markers span six known genes, including the nicotinic acetylcholine receptor subunits *CHRNA5*, *CHRNA3* and *CHRNB4*, the *IREB2* iron-sensing response element, *PSMA4*, which is implicated in DNA repair, and LOC123688, a gene of

Table 1 | Lung cancer risk and rs8034191 genotype

| | Cases* | Controls* | T/C versus T/T genotype | | C/C versus T/T genotype | | Co-dominant model | | P-values | P-heterogeneity |
			OR	95% CI	OR	95% CI	OR	95% CI		
Overall	4,435	7,272	1.21	1.11–1.31	1.77	1.58–2.00	1.30	1.23–1.37	5×10^{-20}	
By study										0.951
Central Europe	1,922	2,520	1.27	1.11–1.44	1.80	1.49–2.18	1.32	1.21–1.45	9×10^{-10}	
Toronto	330	453	1.20	0.85–1.68	1.84	1.14–2.97	1.32	1.05–1.65	0.017	
EPIC	781	1,578	1.18	0.97–1.43	1.68	1.29–2.19	1.27	1.12–1.44	2×10^{-4}	
CARET	764	1,515	1.31	1.08–1.58	1.77	1.34–2.34	1.33	1.16–1.51	2×10^{-5}	
Liverpool	403	814	1.04	0.80–1.34	1.65	1.11–2.44	1.20	1.00–1.44	0.047	
HUNT/ Tromsö	235	392	1.09	0.77–1.54	2.02	1.21–3.37	1.32	1.04–1.68	0.022	

Odds ratio (OR) and 95% confidence interval (CI) for lung cancer comparing heterozygous (T/C) and homozygous (C/C) genotypes of rs8034191 to homozygous (T/T) genotype, overall and separately for each of the six studies. ORs are standardized by age, sex and country. *P*-values are derived from the co-dominant model.
* Subjects with valid call for rs8034191.

unknown function (Fig. 2). It is not possible to identify likely causal alleles or genes based on the differences in the strength of the statistical association because of the strong linkage disequilibrium. However, the nicotinic acetylcholine receptor subunits are strong candidate genes. *CHRNA5* was the only gene found to contain a non-synonymous variant (rs16969968 in exon 5) with strong disease association ($P = 3 \times 10^{-9}$). *CHRNA3* contained a synonymous variant in exon 5 (rs1051730) that was also strongly associated with disease ($P = 5 \times 10^{-9}$); the r^2 between these two variants being 0.99. Although the other markers with a strong disease association either resided in introns or were inter-genic, we cannot exclude the possibility that they could have a biological effect on one or more of the genes from the region. However, other lines of evidence support a possible role for the nicotinic acetylcholine receptor subunit genes.

Nicotinic acetylcholine receptor subunit genes code for proteins that form receptors present in neuronal and other tissues, in particular alveolar epithelial cells, pulmonary neuroendocrine cells, and lung cancer cell lines[2,3], and they bind to nicotine and nicotine derivatives including *N′*-nitrosonornicotine. An association of *CHRNA3* and *CHRNA5* variants with nicotine dependence has been reported[11,13]. The associated markers include the non-synonymous *CHRNA5* SNP, rs16969968, which is one of our markers of lung cancer risk. This SNP introduces a substitution of aspartic acid (D) to asparagine (N) at amino acid position 398 (D398N) of the CHRNA5 protein, located in the central part of the second intracellular loop. Although the function of the second intracellular loop and the possible biological consequences of the D398N alteration remain to be elucidated, this amino acid is highly conserved across species, suggesting that it could have functional importance (Supplementary Fig. 3). A T529A substitution in the second intracellular loop of α4 nAchR, another nicotinic acetylcholine receptor subunit, is known to lead to altered responses to nicotine exposure in the mouse[14].

Within the ARCAGE study (see above), all participants were asked a series of questions relating to tobacco addiction based on the Fagerstrom tolerance questionnaire[15], and we used these to examine whether the chromosome 15q25 locus might be implicated in lung cancer through involvement in tobacco dependence. Two of these questions ('time to first cigarette' and 'numbers of cigarettes per day') have been shown to be particularly strongly associated with nicotine dependence, and responses to both questions result in a 'heaviness of smoking index (HSI)' with a score of between 0 and 6 (ref. 16). We did not observe an association in the ARCAGE controls between rs16969968 and any of the individual Fagerstrom indices of nicotine addiction, or when comparing controls with a HSI of 0 to those with a HSI of 3 or more (Supplementary Table 4). Almost identical patterns were observed for rs8034191 (data not shown). Thus, our data do not support an important role for the locus in nicotine addiction. However, a previous study of a large number of candidate gene markers (4,309 SNPs) identified a possible association between rs16969968 and addiction (uncorrected P-value = 6.4×10^{-4}) using contrasting extreme phenotypes as measured by the Fagerstrom test for nicotine dependence (FTND)[11]. A second study also identified an association between variants in the region of chromosome 15q25 and numbers of cigarettes smoked per day, although it did not assess directly rs16969968[13]. The FTND and HSI measures of nicotine dependence are highly correlated together, and with cigarettes per day[17], and additional studies to clarify the relationship between chromosome 15q25 variants and tobacco dependence are warranted in light of these results.

Our observation of an increased risk with the chromosome 15q25 locus and lung cancer in non-smokers, as well as the lack of an association with smoking-related head and neck cancers, would indicate that the disease mechanism with lung cancer is unlikely to be explained by an association with tobacco addiction. Independent biological data also suggest that nicotinic acetylcholine receptors could be involved in lung cancer through other mechanisms. It has been suggested that *N′*-nitrosonornicotine and nitrosamines may

facilitate neoplastic transformation by stimulating angiogenesis and tumour growth mediated through their interaction with nicotinic acetylcholine receptors[18–20]. The expression of these receptors can also be inhibited by nicotine receptor antagonists, which, if confirmed to be involved in disease aetiology through such a mechanism, implies possible chemoprevention opportunities for lung cancer[5].

No markers outside of those on chromosome 15q25 exceeded the genome-wide significance level for association with lung cancer, although a further 29 had a significance level of $P < 5 \times 10^{-5}$ (Supplementary Table 5). Although most were isolated markers, ten were found to be clustered in a segment of approximately 1 megabase (Mb) on chromosome 6p (28.5–29.5 Mb) within an extended region of high linkage disequilibrium around the major histocompatibility complex. Genotyping of the most significant SNP from the 6p region (rs4324798) in the other five studies provided independent evidence of association ($P = 4 \times 10^{-3}$). In the combined data set, the trend test reached genome-wide significance ($P = 4 \times 10^{-7}$; see Supplementary Fig. 4). The region contains up to 20 documented genes and identification of causal variants is complicated by strong linkage disequilibrium between variants within neighbouring human leukocyte antigen (HLA) and non-HLA genes[21]. Further analyses in multiple diverse populations will be required to confirm this locus and to identify additional lung cancer susceptibility variants. To aid in this, we have made our genome-wide association results available through a publicly accessible website (http://www.ceph.fr/cancer).

METHODS SUMMARY

A detailed description of the component studies can be found in the Supplementary Methods. The genotyping of the IARC central Europe study was conducted using Illumina Sentrix HumanHap300 BeadChip. We excluded variants with a call rate of less than 95% or whose allele distributions deviated strongly from Hardy–Weinberg equilibrium among controls. We also excluded subjects with a completion rate less than 90% or whose reported sex did not match with the inferred sex based on the heterozygosity rate from the X chromosomes. Unexpected duplicates and unexpected first-degree relatives were also excluded from the analysis. Additional quality control measures were applied as described in the Supplementary Methods. Population outliers were detected using STRUCTURE[22] with HapMap subjects as internal controls, and were subsequently excluded from the analysis. Additional analyses for population stratification were undertaken with EIGENSTRAT[23]. Odds ratios (OR) and 95% confidence intervals (CI) were calculated using multivariate unconditional logistic regression models. CEU HapMap SNPs were imputed using MACH (http://www.sph.umich.edu/csg/abecasis/MACH/index.html). Genotyping of additional markers was undertaken with Taqman or Amplifluor assays. Genotyping for all five replication studies was conducted for rs8034191 and rs16969968, and effect estimates from all six lung cancer studies were combined using a fixed-effect model. All P-values are two-sided.

1. Ferlay, J., Bray, F., Pisani, P. & Parkin, M. *GLOBOCAN 2002. IARC CancerBase No 5, version 2.0* (IARC, Lyon, 2004).
2. Minna, J. D. Nicotine exposure and bronchial epithelial cell nicotinic acetylcholine receptor expression in the pathogenesis of lung cancer. *J. Clin. Invest.* **111**, 31–33 (2003).
3. Wang, Y. *et al.* Human bronchial epithelial and endothelial cells express α7 nicotinic acetylcholine receptors. *Mol. Pharmacol.* **60**, 1201–1209 (2001).
4. Schuller, H. M. Nitrosamines as nicotinic receptor ligands. *Life Sci.* **80**, 2274–2280 (2007).
5. Russo, P., Catassi, A., Cesario, A. & Servent, D. Development of novel therapeutic strategies for lung cancer: targeting the cholinergic system. *Curr. Med. Chem.* **13**, 3493–3512 (2006).
6. International Agency for Research on Cancer. Reversal of risk after quitting smoking. *IARC Handbooks of Cancer Prevention* Vol. 11, 15–27 (IARC, Lyon, 2007).
7. International Agency for Research on Cancer. Tobacco smoke and involuntary smoking. *IARC Monographs* Vol. 83, 33–47 (IARC, Lyon, 2004).
8. Coleman, M. P. *et al.* EUROCARE Working Group. EUROCARE-3 summary: cancer survival in Europe at the end of the 20th century. *Ann. Oncol.* **14** (suppl. 5), v128–v149 (2003).
9. Matakidou, A., Eisen, T. & Houlston, R. S. Systematic review of the relationship between family history and lung cancer risk. *Br. J. Cancer* **93**, 825–833 (2005).
10. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* **38**, 659–662 (2006).

11. Saccone, S. F. et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum. Mol. Genet. **16**, 36–49 (2007).
12. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature **449**, 851–861 (2007).
13. Berretini, W. et al. α-5/α-3 nicotinic receptor subunit alleles increase risk for heavy smoking. Mol. Psychiatry advance online publication doi:10.1038/sj.mp.4002154 (29 January 2008).
14. Tritto, T., Stitzel, J. A., Marks, M. J., Romm, E. & Collins, A. C. Variability in response to nicotine in the LSxSS RI strains: potential role of polymorphisms in α4 and α6 nicotinic receptor genes. Pharmacogenetics **12**, 197–208 (2002).
15. Fagerstrom, K. O. & Schneider, N. G. Measuring nicotine dependence: a review of the Fagerstrom Tolerance Questionnaire. J. Behav. Med. **12**, 159–182 (1989).
16. Heatherton, T. F., Kozlowski, L. T., Frecker, R. C. & Fagerstrom, K. O. A Fagerstrom test for nicotine dependence: a revision of the Fagerstrom tolerance questionnaire. Br. J. Addict. **86**, 1119–1127 (1991).
17. Chabrol, H. et al. Comparison of the Heavy Smoking Index and of the Fagerstrom test for nicotine dependence in a sample of 749 cigarette smokers. Addict. Behav. **30**, 1474–1477 (2005).
18. Lam, D. C. et al. Expression of nicotinic acetylcholine receptor subunit genes in non-small-cell lung cancer reveals differences between smokers and nonsmokers. Cancer Res. **67**, 4638–4647 (2007).
19. West, K. A. et al. Rapid Akt activation by nicotine and a tobacco carcinogen modulates the phenotype of normal human airway epithelial cells. J. Clin. Invest. **111**, 81–90 (2003).
20. Dasgupta, P. & Chellappan, S. P. Nicotine-mediated cell proliferation and angiogenesis: new twists to an old story. Cell Cycle **5**, 2324–2328 (2006).
21. De Bakker, P. I. et al. A high resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nature Genet. **38**, 1166–1172 (2006).
22. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**, 1567–1587 (2003).
23. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. **38**, 904–909 (2006).

Author Contributions P.B. and M.L. designed the study. R.J.H., J.D.M., A.B. and H.B. coordinated the preparation and inclusion of all biological samples. R.J.H., J.D.M., V.G. and S.H. undertook the statistical analysis. Bioinformatics analysis was undertaken by F.M., M.F. and S.H., D.Z. and M.D. coordinated the genotyping of the central Europe samples, and J.D.M, R.J.H. and V.G. coordinated the genotyping of the other studies. All other co-authors coordinated the initial recruitment and management of the studies. M.L. obtained financial support for genotyping of the central Europe study, and P.B. and R.J.H. obtained financial support for genotyping of the other studies. P.B. and M.L. drafted the manuscript with substantial contributions from R.J.H. and J.D.M. All authors contributed to the final paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.B. (brennan@iarc.fr).

nature

LETTERS

A variant associated with nicotine dependence, lung cancer and peripheral arterial disease

Thorgeir E. Thorgeirsson[1]*, Frank Geller[1]*, Patrick Sulem[1]*, Thorunn Rafnar[1]*, Anna Wiste[1,2],
Kristinn P. Magnusson[1], Andrei Manolescu[1], Gudmar Thorleifsson[1], Hreinn Stefansson[1], Andres Ingason[1],
Simon N. Stacey[1], Jon T. Bergthorsson[1], Steinunn Thorlacius[1], Julius Gudmundsson[1], Thorlakur Jonsson[1],
Margret Jakobsdottir[1], Jona Saemundsdottir[1], Olof Olafsdottir[1], Larus J. Gudmundsson[1], Gyda Bjornsdottir[1],
Kristleifur Kristjansson[1], Halla Skuladottir[3], Helgi J. Isaksson[4], Tomas Gudbjartsson[5], Gregory T. Jones[8],
Thomas Mueller[9], Anders Gottsäter[10], Andrea Flex[11], Katja K. H. Aben[12,13], Femmie de Vegt[12], Peter F. A. Mulders[14],
Dolores Isla[15], Maria J. Vidal[15], Laura Asin[16], Berta Saez[17], Laura Murillo[18], Thorsteinn Blondal[19],
Halldor Kolbeinsson[6], Jon G. Stefansson[6], Ingunn Hansdottir[20], Valgerdur Runarsdottir[20], Roberto Pola[11,21],
Bengt Lindblad[10], Andre M. van Rij[8], Benjamin Dieplinger[9], Meinhard Haltmayer[9], Jose I. Mayordomo[15,16,17],
Lambertus A. Kiemeney[12,13,14], Stefan E. Matthiasson[22], Hogni Oskarsson[23], Thorarinn Tyrfingsson[20],
Daniel F. Gudbjartsson[1], Jeffrey R. Gulcher[1], Steinn Jonsson[7], Unnur Thorsteinsdottir[1,22], Augustine Kong[1]
& Kari Stefansson[1,22]

Smoking is a leading cause of preventable death, causing about 5 million premature deaths worldwide each year[1,2]. Evidence for genetic influence on smoking behaviour and nicotine dependence (ND)[3–8] has prompted a search for susceptibility genes. Furthermore, assessing the impact of sequence variants on smoking-related diseases is important to public health[9,10]. Smoking is the major risk factor for lung cancer (LC)[11–14] and is one of the main risk factors for peripheral arterial disease (PAD)[15–17]. Here we identify a common variant in the nicotinic acetylcholine receptor gene cluster on chromosome 15q24 with an effect on smoking quantity, ND and the risk of two smoking-related diseases in populations of European descent. The variant has an effect on the number of cigarettes smoked per day in our sample of smokers. The same variant was associated with ND in a previous genome-wide association study that used low-quantity smokers as controls[18,19], and with a similar approach we observe a highly significant association with ND. A comparison of cases of LC and PAD with population controls each showed that the variant confers risk of LC and PAD. The findings provide a case study of a gene–environment interaction[20], highlighting the role of nicotine addiction in the pathology of other serious diseases.

To perform a genome-wide association (GWA) study of smoking quantity (SQ), we used questionnaire data limited to basic questions on smoking behaviour that were available for a large number of lifetime smokers. The GWA scan comprises 10,995 Icelandic smokers who had been assayed with Infinium HumanHap300 SNP chips from Illumina. A set of 306,207 single-nucleotide polymorphisms (SNPs) fulfilling our quality criteria was tested. We focused on cigarette

smoking, with SQ reported as cigarettes per day. All SQ data were clustered into categories (see Supplementary Information) and we refer to them as 'SQ levels'. The SQ levels were 0 (1–10 cigarettes per day), 1 (11–20), 2 (21–30) and 3 (31 or more). Each increment represents an increase in SQ of 10 cigarettes per day. Allele T of the SNP rs1051730 was most strongly associated with SQ, and the association was highly significant ($P = 5 \times 10^{-16}$). The SNP is within the *CHRNA3* gene in a linkage disequilibrium block also containing two other genes, *CHRNA5* and *CHRNB4*, that encode nicotinic acetylcholine receptors (ref. 18). Six other SNPs on chromosome 15q24 passed the threshold of genome-wide significance ($P < 2 \times 10^{-7}$), but they are all correlated with rs1051730 ($r^2 = 0.14$–0.93). After correction for rs1051730, none of these six SNPs showed a P value below 10^{-3} (see Supplementary Table 1). A quantile–quantile plot for the GWA scan (see Supplementary Fig. 1a) shows the observed excess of signals, whereas a quantile–quantile plot after removing 182 markers located within 1 megabase of rs1051730 is consistent with noise (see Supplementary Fig. 1b), illustrating that all of the strongest signals standing out in the first plot are located on chromosome 15q24. An additional 2,950 smokers from Iceland were genotyped for rs1051730, giving a total of 13,945 smokers (Table 1) with a mean variant frequency of 34.7%, which is not significantly different from the frequency of 34.4% observed in 4,203 individuals who were genotyped and who reported never having smoked (odds ratio (OR) 1.01, 95% confidence interval (CI) 0.96–1.07, $P = 0.60$). Indeed, the frequency of the variant in the 3,627 low-quantity smokers (10 or fewer cigarettes per day) is significantly less than that in those who do not smoke (OR 0.83, 95% CI 0.78–0.90, $P = 4.5 \times 10^{-7}$). The increase in

Table 1 | Genotype status and SQ level of 13,945 Icelandic smokers

Parameter	Genotype of rs1051730			Total *n* (frequency)	Frequency of T allele
	GG	GT	TT		
Cigarettes per day (SQ level)					
1 to 10 (0)	1,743	1,558	326	3,627 (0.260)	0.305
11 to 20 (1)	2,727	2,865	810	6,402 (0.459)	0.350
21 to 30 (2)	1,145	1,416	427	2,988 (0.214)	0.380
31 and more (3)	341	448	139	928 (0.067)	0.391
All levels (frequency)	5,956 (0.427)	6,287 (0.451)	1,702 (0.122)	13,945 (1.000)	0.347
Mean SQ level (mean ± s.d.)	1.01 ± 0.85	1.12 ± 0.86	1.22 ± 0.85	1.09 ± 0.86	

frequency between levels varies, and the largest increase (4.5%) is observed between the lowest levels (0 and 1), whereas the increase between the highest levels (2 and 3) is only 1.1%. In the context of a case-control LC study, an additional 523 smokers from Spain and 1,375 smokers from The Netherlands were genotyped. We performed multiple regression analyses of SQ data from the three countries, with adjustment for sex and year of birth (Table 2). Results from Spain and The Netherlands combined gave an estimated increase of 0.074 SQ units ($P = 0.012$) for each copy of the variant, which is not significantly different ($P = 0.45$) from the estimate of 0.098 SQ units ($P = 10^{-18}$) based on the Icelandic data. Combining all results, each copy of the variant was estimated to increase SQ level by 0.095 units ($P = 6 \times 10^{-20}$), which corresponds to about one cigarette per day. A recent GWA study reported an association between SQ and SNP rs6495308 ($P = 6.9 \times 10^{-5}$, correlation between this SNP and rs1051730, $r^2 = 0.18$ in the HapMap project) for about 7,500 individuals from two study groups, as well as association between SQ and rs1317286 ($P = 2.6 \times 10^{-6}$, $r^2 = 0.90$ to rs1051730 in the HapMap project) in a candidate gene study based on 1,740 heavy smokers (more than 25 cigarettes per day) and 6,200 low-quantity smokers (fewer than 5 cigarettes per day)[21].

Sex and year of birth are also strongly associated with SQ (Table 2). However, neither the interaction between variant and sex nor that between variant and year of birth is significant, indicating that the effect of the variant is similar for both sexes and is robust to population-wide changes in smoking habits over time. The phenotypic variance explained by the variant was highest in Iceland, amounting to 0.7%.

Association of the same variant with ND was previously reported in a candidate gene study involving 3,713 SNPs[18]. We assessed the association with ND, defined as a score of 4 or higher on the Fagerstrom Test for Nicotine Dependence (FTND)[22] or endorsement of at least three of the seven Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM-IV) criteria (see Supplementary Information). The variant is associated with ND in Iceland in a subset of 2,394 smokers from the SQ study tested against both 28,455 population controls (OR 1.17, 95% CI 1.10–1.25, $P = 3.3 \times 10^{-6}$) and 3,506 low-quantity smokers (OR 1.40, 95% CI 1.29–1.52, $P = 7 \times 10^{-15}$). The latter OR of 1.40 is comparable to the results of the candidate gene study[18]. They used non-ND smokers as controls (that is, individuals who had smoked but had an FTND score of 0), and reported association with rs1051730 (OR 1.3, $P = 10^{-3}$). rs1051730 is in strong linkage disequilibrium with rs16969968 ($r^2 = 0.90$ in the HapMap project), which was highlighted in the previous study with a similar result (OR 1.3, $P = 6 \times 10^{-4}$)[18].

Dependence on nicotine drives repeated self-administration of nicotine[23–25] and high SQ is a strong sign of ND and one of the criteria for a diagnosis of ND (SQ is included in the FTND scale). This, together with the fact that the ND subjects are part of the SQ study, means that the associations with ND and SQ cannot be considered independent results. Both the FTND and the DSM-IV scales include many items that are not based on SQ, and their total scores are measures of ND severity. In our ND group, positive scores on most items in both scales show a trend towards higher frequency of the variant, as does the total score on both the FTND and DSM-IV scales. Thus, the frequency of the variant increases with addiction severity, and is 46.8% and 43.8% for the highest deciles of FTND and DSM-IV, respectively (see Supplementary Table 2a, b). ND is believed to be the main reason for continued smoking. To explore the frequency of the variant in the context of the ability to quit smoking, we investigated differences between 6,388 current and 6,687 past smokers from the SQ analysis by a logistic regression model adjusting for sex and year of birth. The variant was associated with current smoking with an OR of 1.07 (95% CI 1.01–1.13, $P = 0.015$) (see Supplementary Table 3), and the effect is similar when corrected for SQ (OR 1.06, 95% CI 1.00–1.12, $P = 0.036$), indicating that carriers of the variant are less likely to quit smoking.

Smoking is a major risk factor for many diseases, and we decided to study the effect of the variant on LC and PAD risk directly. The LC study was based on 1,024 cases and 32,244 controls from Iceland, Spain and The Netherlands (Table 3); the PAD study was based on 2,738 cases and 29,964 controls from five caucasian populations (Iceland, New Zealand, Austria, Sweden and Italy) (Table 3). The results for LC and PAD (Table 4) represent the overall effect on LC and PAD including indirect effects through SQ and ND. Significant association was observed with LC for both the Icelandic data (OR 1.27, $P = 4.1 \times 10^{-5}$) and the data for Spain and The Netherlands combined (OR 1.39, $P = 6.6 \times 10^{-5}$). These two estimates are not significantly different from each other ($P = 0.34$), and combining results from all three groups gave an OR of 1.31 (95% CI 1.19–1.44, $P = 1.5 \times 10^{-8}$). There is no significant difference in frequency of the variant between histological types of LC, which is not surprising given the small number of cases per group (see Supplementary Table 4). Association with PAD was found both in the Icelandic data (OR 1.18, $P = 5.3 \times 10^{-5}$) and in the data for the foreign populations combined (OR 1.23, $P = 5.9 \times 10^{-4}$). These two estimates are not significantly different from each other ($P = 0.57$), and combining results from all five groups gave an OR of 1.19 (95% CI 1.12–1.27, $P = 1.4 \times 10^{-7}$).

Table 2 | Multiple regression of SQ level as a function of rs1051730 genotype, sex and year of birth

Study group	*n*	Copies of T allele		Sex (male)		*P* for year of birth (categorical)	*P* for interactions: allele × sex, allele × age
		Estimate (95% CI)	*P*	Estimate (95% CI)	*P*		
Iceland	13,945	0.098 (0.076–0.120)	10^{-18}	0.411 (0.383–0.438)	$<10^{-16}$	$<10^{-16}$	0.53, 0.85
Spain	523	0.061 (−0.059–0.180)	0.32	0.504 (0.290–0.718)	$<10^{-5}$	0.006	0.80, 0.76
The Netherlands	1,375	0.078 (0.012–0.145)	0.021	0.326 (0.225–0.427)	$<10^{-9}$	$<10^{-4}$	0.68, 0.27
Foreign combined	1,898	0.074 (0.016–0.132)	0.012	NA	–	–	–
All combined	15,771	0.095 (0.075–0.115)	6×10^{-20}	NA	–	–	–

Multiple regression of SQ level on allele T, sex and year of birth, giving adjusted values for each explanatory variable adjusting for the others. For the tests of interaction, the interaction terms involving the variant were individually added to the initial model. NA, not available.

639

Table 3 | Demographics, age and phenotype breakdown

Study group	Cases			Controls		
	n	Male/female	Age (years)	n	Male/female	Age (years)
Smoking phenotypes						
Cigarettes per day available						
Iceland	13,945	6,134/7,811	58.7 ± 17.8			
Spain	523	354/169	54.0 ± 16.3			
The Netherlands	1,375	762/613	61.5 ± 10.2			
ND (Iceland)	2,394	800/1,594	48.1 ± 11.0	28,455	12,600/15,855	58.7 ± 21.8
Never smokers (Iceland)	4,203	1,273/2,930	55.4 ± 21.6			
LC						
Iceland	665	346/319	69.7 ± 11.1	28,752	12,174/16,578	56.8 ± 21.5
Spain	269	238/31	64.9 ± 11.6			
The Netherlands	90	71/19	68.5 ± 9.5			
PAD						
Iceland	1,503	926/577	74.2 ± 10.6	28,752	12,174/16,578	56.8 ± 21.5
New Zealand	441	251/189	70.6 ± 9.6	435	248/187	68.2 ± 6.4
Austria	457	322/135	68.4 ± 11.0	403	284/119	67.3 ± 10.7
Sweden	172	92/80	77.5 ± 9.9	140	64/76	67.9 ± 1.5
Italy	165	111/54	73.0 ± 9.3	234	162/72	72.6 ± 6.4

Ages are shown as means ± s.d.

Genotypic ORs for LC, PAD and ND did not deviate significantly from those obtained for the multiplicative model (see Supplementary Table 5), and no significant differences in the ORs between sexes were observed (see Supplementary Table 6).

According to our estimates for Icelandic LC patients, the correlation between SQ and LC is consistent with numbers reported in other studies[26,27]. Combining these estimates with our estimate of the association of the variant with SQ, the expected OR between the variant and LC is only about 1.05 in Iceland (see Supplementary Information), which is well below the direct OR estimate for LC of 1.27 (95% CI 1.13–1.43). A similar indirect estimate for PAD is 1.04, which again is substantially lower than the observed direct estimate of 1.18 (95% CI 1.09–1.27). It is not surprising that the ORs for LC and PAD cannot be explained by the effect of the variant on SQ alone, because the involvement of both SQ and the duration of smoking in LC and PAD was established in previous studies[15,28]. The SQ data for most individuals were derived from a single point in time and cannot be expected to cover all aspects of smoking behaviour affected by the variant and relevant to LC and PAD. An effect on other aspects of smoking behaviour, in particular smoking duration, is likely to account for the observed difference between the indirect and direct estimates of the LC and PAD risks. An alternative possibility is that the variant directly confers risk of LC and PAD, for example by increasing the vulnerability to tobacco smoke.

Thus, we have unequivocally demonstrated a correlation between, on the one hand, a sequence variant in the cluster of genes on chromosome 15 that encode nicotinic acetylcholine receptors and, on the other, SQ and ND. The variant does not influence smoking initiation; however, among smokers, carriers of the variant smoke more than non-carriers and have higher rates of ND. This variant was reported in a previous study of 1,050 ND cases and 879 controls who smoked and had an FTND score of 0 (refs 18, 19) and the authors concluded that the variant contributes to ND (ref. 18). This conclusion is put on firm ground by the highly significant OR of 1.40 ($P = 7 \times 10^{-15}$) for ND compared with low-quantity smokers (ten or fewer cigarettes per day). The direct measurement of the risk of LC and PAD revealed genome-wide significant associations with allelic ORs of 1.31 and 1.19, respectively. This demonstrates that a sequence variant associated with ND, a brain disorder, confers risk of lung and cardiovascular diseases through an effect on behaviour, which is an example of active gene–environment correlation[20] in the pathogenesis of disease. A calculation of the population attributable risk for the variant gives 18% for LC and 10% for PAD. Although these population attributable risks are at best approximate figures given the complex interplay between the variant, smoking, and smoking-related diseases, it is likely that the variant accounts for a substantial fraction of PAD and LC cases and the associated morbidity and mortality.

The results of the study described here show that it is important to keep in mind, while attempting to shed light on the role of nature versus nurture in the pathogenesis of common or complex disease, that variants in the sequence of our genome influence not only how we respond to our environment but also our tendency to seek or avoid environment. The line between nature and nurture is therefore sometimes conspicuously absent.

METHODS SUMMARY

Subjects. Written informed consent was obtained from all subjects in the seven participating populations (Iceland, Spain, The Netherlands, Sweden, Italy, Austria and New Zealand; see Table 3). Inclusion in the study required the availability of genotypes from either GWA studies performed at deCODE

Table 4 | Association of rs1051730 allele T with LC and PAD

Study group	Controls		Cases		OR (95% CI)	P
	n	Frequency	n	Frequency		
LC						
Iceland	28,752	0.342	665	0.398	1.27 (1.13–1.43)	4.1×10^{-5}
Spain	1,474	0.390	269	0.483	1.46 (1.22–1.76)	5.4×10^{-5}
The Netherlands	2,018	0.314	90	0.350	1.18 (0.86–1.61)	0.31
Foreign combined	3,492	–	359	–	1.38 (1.18–1.62)	6.6×10^{-5}
All combined	32,244	–	1,024	–	1.31 (1.19–1.44)	1.5×10^{-8}
PAD						
Iceland	28,752	0.342	1,503	0.379	1.18 (1.09–1.27)	5.3×10^{-5}
New Zealand	435	0.274	441	0.337	1.35 (1.10–1.65)	0.0041
Austria	403	0.352	457	0.395	1.20 (0.99–1.46)	0.068
Sweden	140	0.304	172	0.331	1.14 (0.81–1.60)	0.46
Italy	234	0.378	165	0.412	1.15 (0.86–1.54)	0.33
Foreign combined	1,212	–	1,235	–	1.23 (1.09–1.39)	5.9×10^{-4}
All combined	29,964	–	2,738	–	1.19 (1.12–1.27)	1.4×10^{-7}

genetics or follow-up genotyping of rs1051730 in additional subjects. All subjects were of European descent. For details on recruitment periods and diagnostic criteria used for the ND, LC and PAD groups, see Methods and Supplementary Information.

Association analysis. For the genome-wide study of SQ the significance threshold was set at 2×10^{-7}, which is about 0.05 divided by 306,207, the number of SNPs passing quality control. Regressing SQ level as a quantitative variable on the number of copies of the allele carried (0/1/2), a likelihood ratio χ^2 statistic was used for testing. Evaluation of statistical significance took the relatedness of the individuals into account by dividing the χ^2 statistic by a correction factor based either on the method of genomic control[29] or on a simulation procedure using the known genealogy that we had previously employed[30] (see Supplementary Information). The variant did not correlate with sex or year of birth in the controls; the association analyses were therefore not adjusted for these factors.

ORs were assumed to have log-normal distributions, and the confidence intervals are test-based.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Anonymous. Annual smoking-attributable mortality, years of potential life lost, and productivity losses–United States, 1997–2001. *MMWR Morb. Mortal. Wkly. Rep.* **54**, 625–628 (2005).
2. World Health Organization. Economics of tobacco control. ⟨www.who.int/gb/fctc/PDF/wg1/e1t2.pdf⟩.
3. Carmelli, D., Swan, G. E., Robinette, D. & Fabsitz, R. Genetic influence on smoking—a study of male twins. *N. Engl. J. Med.* **327**, 829–833 (1992).
4. Li, M. D., Cheng, R., Ma, J. Z. & Swan, G. E. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction* **98**, 23–31 (2003).
5. Kendler, K. S. *et al.* A population-based twin study in women of smoking initiation and nicotine dependence. *Psychol. Med.* **29**, 299–308 (1999).
6. Heath, A. C. & Martin, N. G. Genetic models for the natural history of smoking: evidence for a genetic influence on smoking persistence. *Addict. Behav.* **18**, 19–34 (1993).
7. Madden, P. A. *et al.* The genetics of smoking persistence in men and women: a multicultural study. *Behav. Genet.* **29**, 423–431 (1999).
8. Vink, J. M., Willemsen, G., Beem, A. L. & Boomsma, D. I. The Fagerstrom Test for Nicotine Dependence in a Dutch sample of daily smokers and ex-smokers. *Addict. Behav.* **30**, 575–579 (2005).
9. Carlsten, C. & Burke, W. Potential for genetics to promote public health: genetics research on smoking suggests caution about expectations. *J. Am. Med. Assoc.* **296**, 2480–2482 (2006).
10. Bierut, L. J. *et al.* Genetic research and smoking behavior. *J. Am. Med. Assoc.* **297**, 809 (2007).
11. Doll, R. & Hill, A. B. A study of the aetiology of carcinoma of the lung. *BMJ* **2**, 1271–1286 (1952).
12. Doll, R. & Peto, R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. *J. Epidemiol. Community Health* **32**, 303–313 (1978).
13. Wynder, E. L. & Graham, E. A. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. *J. Am. Med. Assoc.* **143**, 329–336 (1950).
14. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA Cancer J. Clin.* **55**, 74–108 (2005).
15. Hirsch, A. T. *et al.* ACC/AHA 2005 Practice Guidelines for the management of patients with peripheral arterial disease (lower extremity, renal, mesenteric, and abdominal aortic): a collaborative report from the American Association for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, and the ACC/AHA Task Force on Practice Guidelines (Writing Committee to Develop Guidelines for the Management of Patients With Peripheral Arterial Disease): endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation; National Heart,
Lung, and Blood Institute; Society for Vascular Nursing; TransAtlantic Inter-Society Consensus; and Vascular Disease Foundation. *Circulation* **113**, e463–e654 (2006).
16. Powell, J. T. *et al.* Risk factors associated with the development of peripheral arterial disease in smokers: a case-control study. *Atherosclerosis* **129**, 41–48 (1997).
17. Price, J. F. *et al.* Relationship between smoking and cardiovascular risk factors in the development of peripheral arterial disease and coronary artery disease: Edinburgh Artery Study. *Eur. Heart J.* **20**, 344–353 (1999).
18. Saccone, S. F. *et al.* Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum. Mol. Genet.* **16**, 36–49 (2007).
19. Bierut, L. J. *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.* **16**, 24–35 (2007).
20. Plomin, R., DeFries, J. C. & Loehlin, J. C. Genotype–environment interaction and correlation in the analysis of human behavior. *Psychol. Bull.* **84**, 309–322 (1977).
21. Berrettini, W. *et al.* α-5/α-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol. Psychiatry* (in the press).
22. Heatherton, T. F., Kozlowski, L. T., Frecker, R. C. & Fagerstrom, K. O. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br. J. Addict.* **86**, 1119–1127 (1991).
23. Rose, J. E. Nicotine addiction and treatment. *Annu. Rev. Med.* **47**, 493–507 (1996).
24. Russell, M. A. The nicotine addiction trap: a 40-year sentence for four cigarettes. *Br. J. Addict.* **85**, 293–300 (1990).
25. Stolerman, I. P. & Jarvis, M. J. The scientific case that nicotine is addictive. *Psychopharmacology (Berl.)* **117**, 2–10 (1995).
26. Haiman, C. A. *et al.* Ethnic and racial differences in the smoking-related risk of lung cancer. *N. Engl. J. Med.* **354**, 333–342 (2006).
27. Stellman, S. D. *et al.* Lung cancer risk in white and black Americans. *Ann. Epidemiol.* **13**, 294–302 (2003).
28. Peto, R. *et al.* Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *Br. Med. J.* **321**, 323–329 (2000).
29. Devlin, B., Bacanu, S. A. & Roeder, K. Genomic control to the extreme. *Nature Genet.* **36**, 1129–1130 (2004).
30. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genet.* **38**, 320–323 (2006).

Author Contributions T.E.T., F.G., P.S., and T.R. contributed equally to this work. T.E.T., F.G., P.S, T.R., A.W., D.F.G., A.K. and K.S. wrote the first draft of the paper. Ha.S., H.J.I., T.G. and S.J. recruited and diagnosed the Icelandic lung cancer patients. S.E.M. recruited and diagnosed the Icelandic peripheral arterial disease patients. T.B., H.K., J.G.S., I.H, V.R., H.O., T.T. and S.J. recruited and diagnosed nicotine addiction subjects. K.K.H.A., F.d.V., P.F.A.M. and L.A.K. recruited and diagnosed the subjects from The Netherlands. D.I., M.J.V., L.A., B.S., L.M. and J.I.M. recruited and diagnosed the Spanish subjects. G.T.J. and A.M.v.R. recruited and diagnosed the subjects from New Zealand. T.M., B.P. and M.H. recruited and diagnosed subjects from Austria. A.G. and B.L. recruited and diagnosed subjects from Sweden. A.F. and R.P. recruited and diagnosed subjects from Italy. A.W., A.I., S.N.S., J.T.B., S.T., J.G., M.J., J.S., O.O. and S.N.S. performed genotyping and experimental work. L.J.G., G.B. and K.K. incorporated phenotypic data into a database and analysed it. T.E.T., F.G., P.S., T.R., A.W., K.P.M., A.M., G.T., D.F.G. and A.K. analysed the data. T.E.T., F.G., P.S., T.R., K.P.M., Hr.S., T.J., J.I.M., L.K., H.O., T.T., J.R.G., S.J., D.G., U.T., A.K. and K.S. planned, supervised and coordinated the work. All authors contributed to the final version of the paper.

Author Information The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.E.T. (thorgeir@decode.is) or K.S. (kari.stefansson@decode.is).

METHODS

Icelandic subjects. For all studies involving Icelandic subjects, the study protocols were approved by the National Bioethics Committee (NBC) and the Data Protection Authority (DPA) of Iceland. The DPA encrypted all personal identifiers associated with information or blood samples with the use of a third-party encryption system[31]. In all, the Icelandic study involves 10,995 subjects with information on SQ available in the GWA, an additional 2,950 subjects with information on SQ, and 4,203 never-smokers. In the studies of LC and PAD, 665 and 1,503 patients, respectively, and 28,752 population controls were used (for details see Table 3).

Smoking. All Icelandic subjects in the study of smoking-related phenotypes, including Icelandic population controls, were originally recruited for different genetic studies conducted over 11 years (1996–2007) at deCODE Genetics, and information on the number of cigarettes smoked per day was available from questionnaires. The information on cigarettes smoked per day was categorized into SQ levels and used as a quantitative variable. Detailed information on SQ was also available for the foreign LC populations (Supplementary Information), but not for the foreign PAD populations.

Nicotine dependence. For a subset of the Icelandic smokers, information on the criteria used to diagnose ND was available from ongoing studies of ND and anxiety/depression[32]. We excluded individuals with diagnoses of other substance dependence or abuse, giving a total of 2,394 ND subjects. A score of 4 or higher on the FTND[22], or endorsement of three or more DSM criteria, were used to assign affected status for ND. Additional information on the Icelandic smoking and ND study group is available in the Supplementary Information.

Lung cancer. Iceland: recruitment was initiated in the year 1998 with a nationwide list from the Icelandic Cancer Registry (ICR). About 1,265 LC patients were alive during the period of recruitment, and 665 participated in the project. Information in the ICR includes year and age at diagnosis, year of death, SNOMED (Systematized Nomenclature of Medicine) code and ICD-10 (International Statistical Classification of Diseases and Related Health Problems, 10th revision) classification. Histological and cytological verification was available for 647 cases; the remaining 18 cases were diagnosed clinically.

The Netherlands: the 90 patients and 2,018 controls were identified retrospectively through three different ongoing studies on genetic risk factors of disease. All three study protocols were approved by the Institutional Review Board of the Radboud University Nijmegen Medical Centre.

Spain: patients were recruited from the Oncology Department of Zaragoza Hospital, from June 2006 to June 2007, and of 330 patients who were invited to participate, 292 enrolled (88%). Clinical information including age at onset and histology were collected from medical records. The 1,474 control individuals were approached at Zaragoza University Hospital. Study protocols were approved by the Institutional Review Board of Zaragoza University Hospital.

PAD. Iceland: patients have been recruited over the past nine years, as part of a genetic study at deCODE, from a registry of individuals diagnosed with PAD at the major hospital in Reykjavik, the Landspitali University Hospital, during the years 1983–2006. Diagnosis was confirmed by vascular imaging or segmental pressure measurements.

Austria: patients and controls were recruited through the Linz Peripheral Arterial Disease (LIPAD) study during 2000 to 2002, at the Department of Surgery, St John of God Hospital. Of the patients admitted for evaluation of suspected or definite PAD, all patients with chronic atherosclerotic occlusive disease of the lower extremities associated with typical symptoms—such as claudication or leg pain on exertion, rest pain, or minor or major tissue loss—were included on the basis of the final clinical diagnosis established by attending vascular surgeons. The diagnosis was verified by interview, physical examination, noninvasive techniques, and angiography[33]. All control subjects were patients at the same hospital and fulfilled the following criteria: no clinical indication of PAD by history and physical examination, and systolic brachial blood pressure equal to or less than the blood pressure in each of the right and left anterior tibial and posterior tibial arteries (that is, ankle brachial index ≥ 1.0)[33]. Smoking status was assessed as described in ref. 34.

Sweden: patients and controls were recruited at the Department of Vascular Diseases at Malmö University Hospital, a single referral centre for all patients with critical limb ischaemia in the three southernmost health-care districts in Sweden (723,750 inhabitants in 2001). The diagnosis of critical limb ischaemia was made in accordance with Trans-Atlantic Inter-Society Consensus scientific criteria[35] of ulceration, gangrene, or rest pain caused by PAD proved by ankle pressure (less than 50 to 70 mmHg), reduced toe pressure (less than 30 to 50 mmHg) or reduced transcutaneous oxygen tension. Diagnosis was confirmed by an experienced vascular surgery consultant. The control group consisted of healthy individuals without symptomatic PAD included in a health screening programme for a preventive medicine project[36].

Italy: patients and controls were recruited from subjects admitted to the Department of Medicine of the A. Gemelli University hospital of Rome, from 2000 to 2001. Inclusion criteria for the PAD group were Caucasian origin and presence of PAD, diagnosed in accordance with established criteria[37]. All patients had an ankle/arm pressure index lower than 0.8 and were at Fontaine's stage II, with intermittent claudication and no rest pain or trophic lesions. Inclusion criteria for the control group were caucasian origin, absence of PAD and CAD and no relationship to cases. Additional exclusion criteria from the study were tumours, chronic inflammatory diseases, and autoimmune diseases[38].

New Zealand: patients were recruited from the Otago–Southland region, and PAD was confirmed by an ankle brachial index of less than 0.7, pulse volume recordings and angiography/ultrasound imaging. The control group consisted of elderly individuals with no history of vascular disease from the same geographical region. Controls were asymptomatic for PAD and had ankle brachial indexes of more than 1. An abdominal ultrasound scan excluded concurrent abdominal aortic aneurysm from both the PAD and control groups, and Anglo-European ancestry was required for inclusion.

Genotyping. All 10,995 samples in the GWA study of SQ were genotyped with genotyping systems and specialized software (Human Hap300 and Human Hap300-duo+ Bead Arrays; Illumina)[39]. rs1051730 was genotyped with a Centaurus assay (Nanogen) for 8,566 Icelandic samples and all samples in the foreign study groups. Information on the genotyping and quality control is given in the Supplementary Information.

31. Gulcher, J. R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8**, 739–742 (2000).
32. Thorgeirsson, T. E. *et al.* Anxiety with panic disorder linked to chromosome 9q in Iceland. *Am. J. Hum. Genet.* **72**, 1221–1230 (2003).
33. Mueller, T. *et al.* Factor V Leiden, prothrombin G20210A, and methylenetetrahydrofolate reductase C677T mutations are not associated with chronic limb ischemia: the Linz Peripheral Arterial Disease (LIPAD) study. *J. Vasc. Surg.* **41**, 808–815 (2005).
34. Rutherford, R. B. *et al.* Recommended standards for reports dealing with lower extremity ischemia: revised version. *J. Vasc. Surg.* **26**, 517–538 (1997).
35. Dormandy, J. A. & Rutherford, R. B. Management of peripheral arterial disease (PAD). TASC Working Group. TransAtlantic Inter-Society Consensus (TASC). *J. Vasc. Surg.* **31**, S1–S296 (2000).
36. Barani, J., Nilsson, J. A., Mattiasson, I., Lindblad, B. & Gottsater, A. Inflammatory mediators are associated with 1-year mortality in critical limb ischemia. *J. Vasc. Surg.* **42**, 75–80 (2005).
37. Anonymous.. Suggested standards for reports dealing with lower extremity ischemia. Prepared by the Ad Hoc Committee on Reporting Standards, Society for Vascular Surgery/North American Chapter, International Society for Cardiovascular Surgery. *J. Vasc. Surg.* **4**, 80–94 (1986).
38. Flex, A. *et al.* The −174 G/C polymorphism of the interleukin-6 gene promoter is associated with peripheral artery occlusive disease. *Eur. J. Vasc. Endovasc. Surg.* **24**, 264–268 (2002).
39. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* **38**, 659–662 (2006).

LETTERS

A neural representation of depth from motion parallax in macaque visual cortex

Jacob W. Nadler[1], Dora E. Angelaki[1] & Gregory C. DeAngelis[1,2]

Perception of depth is a fundamental challenge for the visual system, particularly for observers moving through their environment. The brain makes use of multiple visual cues to reconstruct the three-dimensional structure of a scene. One potent cue, motion parallax, frequently arises during translation of the observer because the images of objects at different distances move across the retina with different velocities. Human psychophysical studies have demonstrated that motion parallax can be a powerful depth cue[1–5], and motion parallax seems to be heavily exploited by animal species that lack highly developed binocular vision[6–8]. However, little is known about the neural mechanisms that underlie this capacity. Here we show, by using a virtual-reality system to translate macaque monkeys (*Macaca mulatta*) while they viewed motion parallax displays that simulated objects at different depths, that many neurons in the middle temporal area (area MT) signal the sign of depth (near versus far) from motion parallax in the absence of other depth cues. To achieve this, neurons must combine visual motion with extra-retinal (non-visual) signals related to the animal's movement. Our findings suggest a new neural substrate for depth perception and demonstrate a robust interaction of visual and non-visual cues in area MT. Combined with previous studies that implicate area MT in depth perception based on binocular disparities[9–12], our results suggest that area MT contains a more general representation of three-dimensional space that makes use of multiple cues.

Humans can make precise judgments of depth on the basis of motion parallax, the relative retinal image motion between objects at different distances[1–5]. However, motion parallax alone is not sufficient to specify the sign of depth; that is, whether an object is near or far relative to the plane of fixation[13–15]. Rather, the direction of image motion relative to observer motion is crucial to specifying depth-sign (Fig. 1). Objects located nearer than the plane of fixation sweep across the visual field in a direction opposite to head translation (Fig. 1, black arrows). In contrast, objects located farther than the plane of fixation move in the same direction as the head (Fig. 1, white arrows). Thus, image motion in a particular direction (for example, leftwards) can be associated with either a near object (Fig. 1a) or a far one (Fig. 1b). Under some conditions, the brain could interpret such ambiguous visual motion by using other cues such as occlusion, size or perspective. However, to compute depth-sign in the absence of these pictorial cues, the visual system needs access to extra-retinal signals related to observer movement. We exploited this fact to probe for a neural correlate of depth from motion parallax.

We performed extracellular microelectrode recordings in area MT of two macaque monkeys that were trained to maintain visual fixation on a world-fixed target while being translated by a motion platform. While tracking the target, the animal viewed a display in which a circular patch of random dots was placed over the neuron's visual

receptive field. In each trial, motion of the dots was computed to accurately simulate a surface placed at one of nine depths, which are expressed in terms of their equivalent binocular disparities (Fig. 1c and Supplementary Fig. 1). Stimuli were viewed monocularly, and the random-dot stimulus was scaled such that the retinal image maintained a constant size, retinotopic location and dot density, independently of simulated depth (Fig. 1c). By eliminating all pictorial depth cues, we forced our visual stimulus to be depth-sign ambiguous.

Our experiment consisted of two main stimulus conditions that were randomly interleaved. In the Motion Parallax (MP) condition,
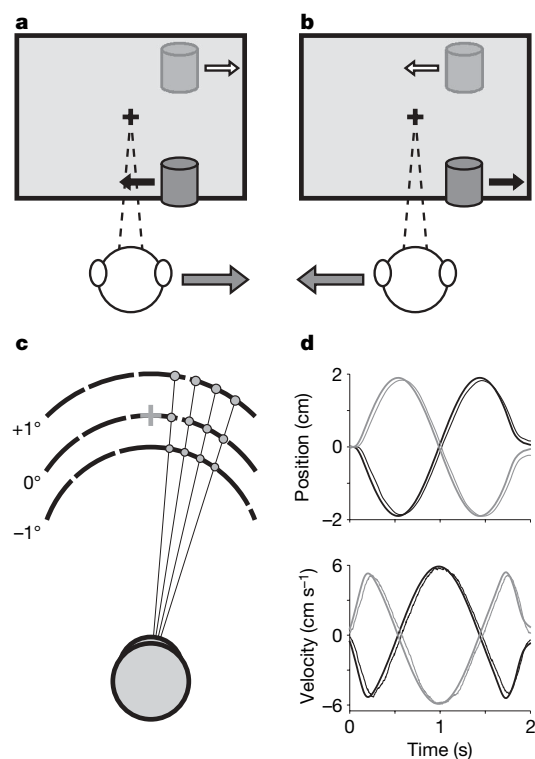

Figure 1 | Schematic illustration of motion parallax and stimulus design. a, As the head moves to the right, the image of a near object moves leftwards, whereas the image of a far object moves rightwards. **b**, The opposite occurs during leftward head movement. Without pictorial depth cues, an extra-retinal signal is needed to determine depth sign. **c**, Random-dot stimuli were scaled so that size and density were identical across simulated depths. Three depths—far ($+1°$), near ($-1°$) and zero—are illustrated. **d**, Thick black and grey curves represent the motion trajectory for two possible starting phases. Thin curves represent average eye position and velocity traces for a single session, in equivalent stimulus units.

[1]Department of Anatomy and Neurobiology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [2]Department of Brain and Cognitive Sciences, University of Rochester, New York 14627, USA.

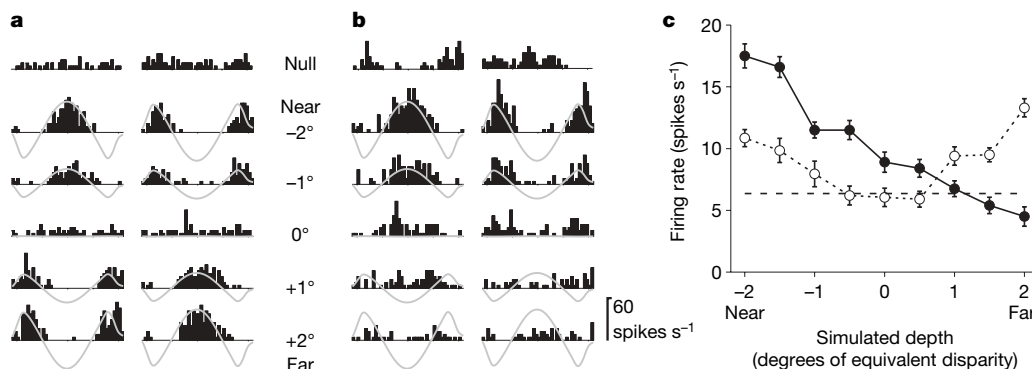**Figure 2 | A neuron selective for depth from motion parallax. a**, Responses in the Retinal Motion (RM) condition for five of nine simulated depths tested. One column of peri-stimulus time histograms is shown for each starting phase of motion. Grey traces represent retinal image velocity, with peaks representing motion in the neuron's preferred direction. 'Null' responses were obtained when no random dots were presented. **b**, The

Motion Parallax (MP) condition. Responses to near stimuli are accentuated, whereas responses to far stimuli are suppressed relative to the RM condition. **c**, Depth-tuning curves for the RM (open symbols, DSDI =0.15) and MP (filled symbols, DSDI = −0.73) conditions. Error bars represent s.e.m.; the dashed horizontal line indicates average spontaneous activity.

the animal was translated sinusoidally at 0.5 Hz (Fig. 1d) while a three-dimensional graphics engine performed the necessary projections to render the fixation point and random-dot surface as stationary world-fixed objects. During translation, the animal was required to make compensatory smooth eye movements to track the fixation point (Fig. 1d and Supplementary Fig. 2). In the Retinal Motion (RM) condition, we replicated the visual image seen during the MP condition. However, the animal remained stationary and therefore did not have to make any eye movements. Because the retinal stimulation was the same in both conditions, any differences in neural response should be due to the action of extra-retinal signals.

Figure 2 illustrates results obtained from a single neuron. Peri-stimulus time histograms are shown for the RM and MP conditions (Fig. 2a, b), with responses grouped (into columns) according to the phase of the real or simulated movement of the observer. Note that near and far simulated depths have opposite retinal velocity profiles (grey traces). In the RM condition (Fig. 2a), responses of this directionally selective MT neuron follow the retinal velocity of the stimulus such that near and far simulated depths of the same magnitude (for example −2° versus +2°) elicit very similar responses. Given the amplitude and speed of the head movements, resulting retinal velocities were slow (ranging from about $0° s^{-1}$ for a simulated depth of 0° up to $\pm 5° s^{-1}$ for the most extreme simulated depths of $\pm 2°$; Supplementary Fig. 1). Because this neuron preferred fast speeds, simulated depths closer to the plane of fixation evoked less activity. When average firing rate is plotted as a function of simulated depth, the neuron shows a depth-tuning curve that is symmetrical around 0° equivalent disparity (Fig. 2c, open symbols). A strikingly different pattern of response occurred in the MP condition (Fig. 2b). Responses to near stimuli (for example −2°) were enhanced, whereas responses to far stimuli (for example +2°) were suppressed relative to the RM condition. As a result, the depth-tuning curve is strongly asymmetric in the MP condition, with response declining monotonically from near to far (Fig. 2c, filled symbols). Because the retinal motion stimulus is the same for the two conditions, this selectivity for depth sign must arise from extra-retinal signals.

Depth-tuning curves for six additional cells are shown in Fig. 3. Three of these example neurons prefer near depths (left) in the MP condition, whereas the other three prefer far depths (right). Note that the tuning curves are generally monotonic within the range tested. Indeed, we found that only 28% of MT neurons with significant depth-sign selectivity had tuning curves that deviated significantly from monotonicity in the MP condition (see Methods), including the example cell in Fig. 3a. This tendency towards monotonic tuning is in contrast with the peaked tuning curves that many MT neurons

show in response to binocular disparities[16,17], but this does not preclude these cells from participating in a representation of depth based on multiple cues.

Depth-sign selectivity was quantified with a depth-sign discrimination index (DSDI; see Methods). Neurons with strong preferences for near and far depths will have DSDIs approaching −1 and +1, respectively. All neurons with well-isolated action potentials were studied ($n = 144$), although some responded poorly at all depths because the range of stimulus speeds was not well matched to their
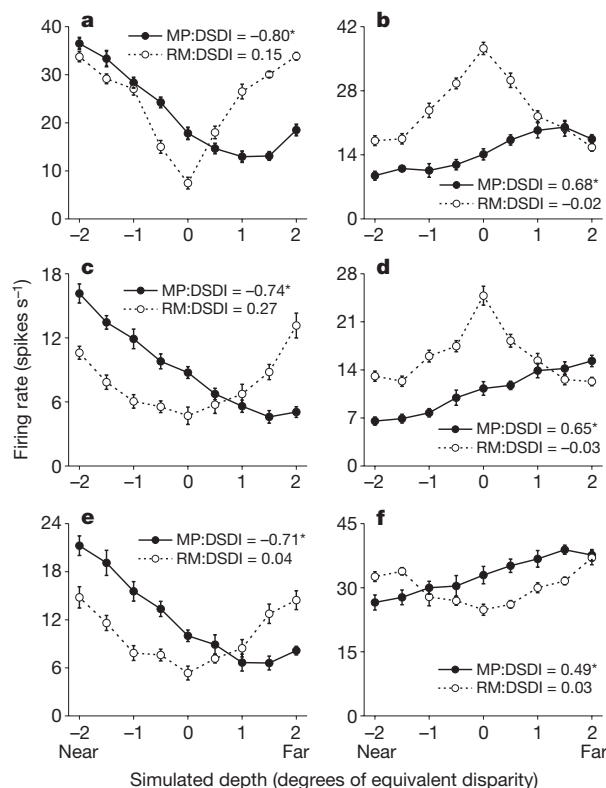

Figure 3 | Tuning curves from six additional MT cells. Filled and open symbols show firing rates (means ± s.e.m.) in the MP and RM conditions, respectively. Asterisks denote significant DSDIs ($P < 0.05$). **a, c, e**, Neurons that prefer near stimuli in the MP condition. Speed preferences are $16° s^{-1}$, $27° s^{-1}$ and $17° s^{-1}$, respectively. **b, d, f**, Neurons that prefer far stimuli. Cells in **b** and **d** prefer slow speeds ($1.0° s^{-1}$ and $0.0° s^{-1}$, respectively) and thus have RM tuning curves that are symmetrically peaked. The neuron in **f** has a moderate speed preference ($5.2° s^{-1}$).

speed preference. In the MP condition (Fig. 4a, top histogram), DSDIs are spread across a wide range, and roughly two-thirds of the neurons (100/144) show significant selectivity ($P < 0.05$, permutation test). Neurons that prefer near depths (66/100) were significantly more common than neurons that prefer far depths (34/100) ($P = 0.0014$, χ^2 test). This held true for both animals individually ($P < 0.04$, χ^2 tests). By contrast, the range of DSDIs in the RM condition is much narrower ($P < 0.0001$, Levene's F-test), and far fewer cells are significantly tuned (17/144). Although this fraction (12%) is larger than expected by chance (5%), this can be explained by a motion parallax equivalent of vertical disparity (Supplementary Fig. 3). Overall, the average absolute value of the DSDI in the MP condition (0.41 ± 0.02; mean $\pm$ s.e.m.) was significantly larger ($P < 0.001$, paired t-test) than that in the RM condition (0.15 ± 0.01). This difference indicates that many MT neurons are modulated by extra-retinal signals to generate selectivity for depth sign.

Retinal image motion is identical in the MP and RM conditions, provided that the monkey accurately tracks the fixation target during platform motion. We must therefore exclude the possibility that differences in selectivity between the MP and RM conditions result from inaccurate eye pursuit. Pursuit was generally good and saccades were rare (Fig. 1d and Supplementary Fig. 2). However, both animals tended to slightly under-pursue the fixation target by about 4% (average pursuit gain 0.963). A careful analysis revealed that there was no significant correlation between the accuracy of pursuit and DSDIs measured in the MP condition (Supplementary Fig. 4). In addition, we corrected DSDIs for imperfect pursuit and found only small changes that would not alter our conclusions (Supplementary Fig. 5). Depth-sign selectivity in area MT therefore cannot be explained by inaccuracies in eye movements.

Figure 4b reveals a significant negative correlation ($r = -0.57$, $P < 0.0001$, Spearman rank correlation) between depth-sign preference in the MP condition and preferred retinal speed. MT neurons that prefer far depths tended to prefer slow speeds, whereas near-preferring neurons were tuned to a broad range of speeds (except for speeds very close to zero). This correlation may reflect an adaptation to the ecological constraint that the range of retinal image speeds (due to observer motion) is larger for near objects than for far objects. Under the conditions of our experiment, any object nearer than one-half the viewing distance will have a retinal speed greater than that of all possible far objects. Alternatively, this correlation may simply be a by-product of the mechanism by which extra-retinal signals interact with visual motion in area MT. In either case, the correlation between DSDI and speed preference probably explains the higher proportion of near-preferring neurons in our sample. Neurons with speed preferences above $5° s^{-1}$ tended to be near-preferring, and most neurons in area MT have speed preferences that exceed this value[18,19]. A similar predominance of near-preferring neurons has been reported for tuning to binocular disparity in area MT[17], although disparity preferences were not found to be strongly correlated with speed preferences[16].

It is sensible that depth from motion parallax would be represented in neural circuits that are sensitive to visual motion and that have a function in coding three-dimensional scene structure. Area MT is well known as a motion-processing area[20]. In addition, most MT neurons are tuned for depth defined by binocular disparities[16,17], and area MT has been causally linked to some forms of depth perception[9,11]. Motion and disparity cues are known to interact in area MT to disambiguate relative image motion that results from three-dimensional object structure[21–23]. Our findings show that, even in the absence of disparity cues, MT neurons can use extra-retinal signals to disambiguate retinal image motion and compute depth. This substantially enhances the flexibility of depth signals in area MT. Neurons that are selective for both disparity and motion parallax should permit a more robust representation of depth, and they may mediate improvements in depth perception seen when disparity and motion parallax cues are presented together[3,5].

It is unclear whether selectivity for depth from motion parallax emerges in area MT, or whether it arises earlier in the visual pathways. Neurons that are sensitive to relative image motion have been described in primary visual cortex of monkeys[24], but this does not necessarily imply a role in computing depth from motion parallax. We have demonstrated selectivity for depth sign by using ambiguous retinal image motion and showing that an extra-retinal signal overcomes this ambiguity. The strength of the extra-retinal influences in our data may seem surprising in view of previous work. MT neurons are not thought to be strongly modulated by pursuit eye movements[25], and we are not aware of any published evidence of substantial vestibular signals in area MT. Responses of MT neurons have been shown to be modulated by eye position[26], but such 'gain fields' cannot explain our results because the same variation in eye position has opposite modulatory effects on responses to near and far stimuli (Fig. 2b).

What might be the mechanism that generates depth-sign selectivity in area MT? One possibility is that an extra-retinal signal related to head or eye movement simply sums with responses to visual motion, thus enhancing responses to one depth sign (for example, near) and suppressing responses to the opposite depth sign[27]. Our
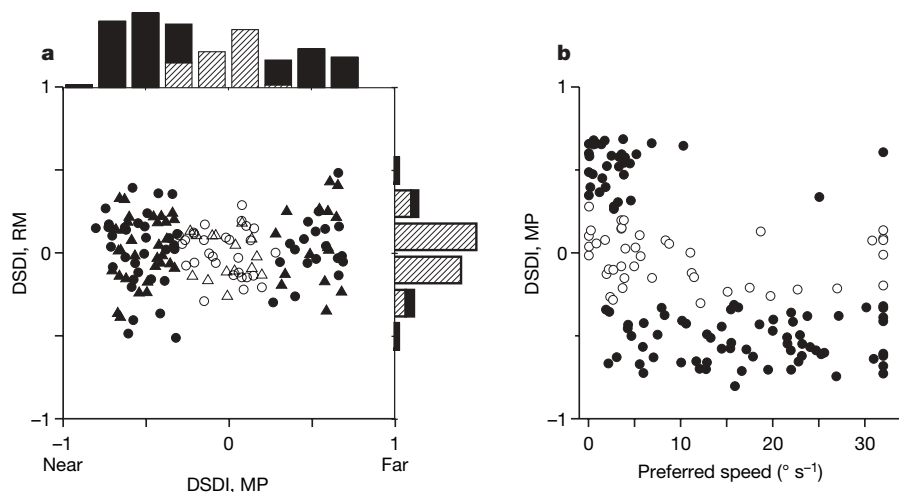

Figure 4 | Many MT neurons are selective for depth defined by motion parallax. a, DSDIs for the RM condition are plotted against those for the MP condition ($n = 144$ neurons from two monkeys: circles, M1; triangles, M2). In the marginal histograms, solid bars represent DSDIs significantly different from zero ($P < 0.05$). **b**, DSDIs in the MP condition are correlated with speed preferences ($r = -0.57$, $P < 0.0001$, Spearman rank correlation). Individual cells are coded for significance of the DSDI in the MP condition (filled, $P < 0.05$; open, $P > 0.05$).

data suggest that the mechanism is more complicated. In 'null' trials containing no visual stimulus in the receptive field, 75% of MT neurons still show significant response modulations during eye/head motion (see, for example, Fig. 2b, top). However, these modulations generally fail to predict, in both sign and magnitude, the difference in activity between the MP and RM conditions (Supplementary Fig. 6). For the neuron of Fig. 2, if the response modulations seen in the null trials (Fig. 2b, top) simply added to responses in the RM condition, the neuron should prefer far depths. Instead it prefers near stimuli (see also Supplementary Fig. 6a). Thus, the interaction between visual motion and extra-retinal signals seems to be complex and nonlinear, and requires further study.

We have shown that single neurons in area MT carry reliable information about depth sign from motion parallax. Although the existence of motion parallax as a potent cue for depth perception has long been established[1,4], our findings provide evidence for a neural substrate for this perceptual capacity. We cannot yet prove that these signals in area MT are used by the monkey to perceive depth (they might reflect extra-retinal signals used for another purpose), but our findings enable a direct causal test in trained animals. Such proof notwithstanding, our results establish a new potential neural mechanism for processing depth information and suggest that area MT may be involved in integrating multiple cues to depth.

METHODS SUMMARY

We recorded extracellular single-unit activity from area MT by using tungsten microelectrodes (FHC) in two adult rhesus monkeys (*Macaca mulatta*). A custom-made virtual-reality system[28] was used to provide stimuli consisting of sinusoidal translation and/or visual motion. Animals were trained to maintain fixation on a visual target during translation of the motion platform. Custom-written OpenGL software was used to generate visual stimuli that depicted a random-dot surface at one of several possible depths in a virtual environment. The visual stimulus was viewed monocularly by the animal, and all pictorial depth cues were removed from the stimulus to render the visual motion ambiguous with respect to depth sign. Thus, to compute depth sign based on motion parallax, neurons needed to combine visual motion with extraretinal signals generated by physical translation of the animal (for example, vestibular or eye movement signals).

Our experimental design compared neural responses in two conditions: a Motion Parallax (MP) condition in which the combination of physical translation and visual motion specified depth unambiguously, and a Retinal Motion (RM) control condition in which the same visual motion stimulus was presented in the absence of extra-retinal signals, thus rendering it ambiguous with respect to depth sign. Neuronal responses were measured as mean firing rates, and the significance of depth-sign selectivity was assessed with permutation tests. Eye position data were filtered (200 Hz lowpass) and analysed to quantify the accuracy of pursuit eye movements.

All procedures were approved by the Institutional Animal Care and Use Committee at Washington University and were in accordance with National Institutes of Health guidelines.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Rogers, B. & Graham, M. Motion parallax as an independent cue for depth perception. *Perception* 8, 125–134 (1979).
2. Rogers, B. J. Motion parallax and other dynamic cues for depth in humans. *Rev. Oculomot. Res.* 5, 119–137 (1993).
3. Rogers, B. J. & Collett, T. S. The appearance of surfaces specified by motion parallax and binocular disparity. *Q. J. Exp. Psychol. A* 41, 697–717 (1989).
4. Rogers, B. J. & Graham, M. E. Similarities between motion parallax and stereopsis in human depth perception. *Vision Res.* 22, 261–270 (1982).
5. Bradshaw, M. F. & Rogers, B. J. The interaction of binocular disparity and motion parallax in the computation of depth. *Vision Res.* 36, 3457–3468 (1996).
6. Ellard, C. G., Goodale, M. A. & Timney, B. Distance estimation in the Mongolian gerbil: the role of dynamic depth cues. *Behav. Brain Res.* 14, 29–39 (1984).
7. van der Willigen, R. F., Frost, B. J. & Wagner, H. Depth generalization from stereo to motion parallax in the owl. *J. Comp. Physiol. A* 187, 997–1007 (2002).
8. Kral, K. Behavioural–analytical studies of the role of head movements in depth perception in insects, birds and mammals. *Behav. Processes* 64, 1–12 (2003).
9. DeAngelis, G. C., Cumming, B. G. & Newsome, W. T. Cortical area MT and the perception of stereoscopic depth. *Nature* 394, 677–680 (1998).
10. Uka, T. & DeAngelis, G. C. Contribution of middle temporal area to coarse depth discrimination: comparison of neuronal and psychophysical sensitivity. *J. Neurosci.* 23, 3515–3530 (2003).
11. Uka, T. & DeAngelis, G. C. Linking neural representation to function in stereoscopic depth perception: roles of the middle temporal area in coarse versus fine disparity discrimination. *J. Neurosci.* 26, 6791–6802 (2006).
12. Uka, T. & DeAngelis, G. C. Contribution of area MT to stereoscopic depth perception: choice-related response modulations reflect task strategy. *Neuron* 42, 297–310 (2004).
13. Farber, J. M. & McConkie, A. B. Optical motions as information for unsigned depth. *J. Exp. Psychol. Hum. Percept. Perform.* 5, 494–500 (1979).
14. Rogers, S. & Rogers, B. J. Visual and nonvisual information disambiguate surfaces specified by motion parallax. *Percept. Psychophys.* 52, 446–452 (1992).
15. Nawrot, M. Eye movements provide the extra-retinal signal required for the perception of depth from motion parallax. *Vision Res.* 43, 1553–1562 (2003a).
16. DeAngelis, G. C. & Uka, T. Coding of horizontal disparity and velocity by MT neurons in the alert macaque. *J. Neurophysiol.* 89, 1094–1111 (2003).
17. Maunsell, J. H. & Van Essen, D. C. Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity. *J. Neurophysiol.* 49, 1148–1167 (1983b).
18. Maunsell, J. H. & Van Essen, D. C. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *J. Neurophysiol.* 49, 1127–1147 (1983a).
19. Nover, H., Anderson, C. H. & DeAngelis, G. C. A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discriminiation performance. *J. Neurosci.* 25, 10049–10060 (2005).
20. Born, R. T. & Bradley, D. C. Structure and function of visual area MT. *Annu. Rev. Neurosci.* 28, 157–189 (2005).
21. Bradley, D. C., Qian, N. & Andersen, R. A. Integration of motion and stereopsis in middle temporal cortical area of macaques. *Nature* 373, 609–611 (1995).
22. Bradley, D. C., Chang, G. C. & Andersen, R. A. Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* 392, 714–717 (1998).
23. Dodd, J. V., Krug, K., Cumming, B. G. & Parker, A. J. Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *J. Neurosci.* 21, 4809–4821 (2001).
24. Cao, A. & Schiller, P. H. Neural responses to relative speed in the primary visual cortex of rhesus monkey. *Vis. Neurosci.* 20, 77–84 (2003).
25. Newsome, W. T., Wurtz, R. H. & Komatsu, H. Relation of cortical areas MT and MST to pursuit eye movements. II. Differentiation of retinal from extraretinal inputs. *J. Neurophysiol.* 60, 604–620 (1988).
26. Bremmer, F., Ilg, U. J., Thiele, A., Distler, C. & Hoffmann, K. P. Eye position effects in monkey cortex. I. Visual and pursuit-related activity in extrastriate areas MT and MST. *J. Neurophysiol.* 77, 944–961 (1997).
27. Nawrot, M. & Joyce, L. The pursuit theory of motion parallax. *Vision Res.* 46, 4709–4725 (2006).
28. Gu, Y., Watkins, P. V., Angelaki, D. E. & DeAngelis, G. C. Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *J. Neurosci.* 26, 73–85 (2006).

Author Contributions J.W.N. and G.C.D. designed the stimuli; J.W.N. collected the data; J.W.N., D.E.A. and G.C.D. refined the analysis and presentation of the data; J.W.N., D.E.A. and G.C.D. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.C.D. (gdeangelis@cvs.rochester.edu).

METHODS

Subjects and task. Two male monkeys (*Macaca mulatta*) were prepared for neurophysiological experiments as described in detail elsewhere[28]. Monkeys were head restrained and seated in front of a 60 cm × 60 cm tangent screen that subtended roughly 90° × 90° of visual angle at the viewing distance of 32 cm. The display screen and stereoscopic projector (Christie Digital Mirage 2000) were mounted on a motion platform with six degrees of freedom (6DOF2000E; MOOG) that allowed us to translate the animal along any direction in the frontoparallel plane (fore–aft movements were not used here). Platform movements and visual stimuli were controlled by computer at 60 Hz, and the measured transfer function of the system (verified by accelerometer measurements) allowed us to record the platform's position accurately at all times[28,29]. A room-mounted laser was used to ensure that platform motion was precisely synchronized with the video display (to within about 1 ms). Additional details are available in a previous publication[28].

For each isolated neuron, we first obtained quantitative measurements of direction tuning and receptive field location, as described previously[30], and we used those measurements to specify the location, size and direction axis of the random-dot stimulus. To measure speed tuning, each neuron was tested with random-dot patterns that had speeds of 0, 0.5, 1, 2, 4, 8, 16 and $32°\,s^{-1}$. Preferred speeds were obtained from the peak of a fitted Gamma function[19]. For the main experimental conditions, the OpenGL graphics library was used to render a fixation target and a random-dot surface in a world-fixed virtual environment (Fig. 1c). The random-dot surface was a portion of a cylinder (analogous to the geometric horopter) that was oriented perpendicular to the axis of translation (Supplementary Fig. 1). At 500 ms after the monkey had achieved fixation, the random-dot stimulus was presented in the neuron's receptive field. In each trial, one of nine simulated depths was chosen pseudo-randomly from the following set of equivalent disparities: 0.0°, ±0.5°, ±1.0°, ±1.5° and ±2.0° (depth from motion parallax is commonly expressed in units of the equivalent binocular disparity[4]). 'Null' trials, in which no stimulus was presented over the receptive field, were also interleaved. Stimuli were viewed monocularly and the random-dot patch was positioned and scaled so as to eliminate all other cues to depth. The random-dot surface was transparent when it overlapped the fixation target, so that no occlusion cues were present (see Supplementary Fig. 1 for additional details of stimulus generation). For almost all cells, visual stimuli were presented to the eye contralateral to the recording hemisphere.

Our experiment consisted of two stimulus conditions. In the Motion Parallax (MP) condition, the animal was translated through one cycle of a 0.5-Hz sinusoid having a total displacement of 4 cm, which is slightly more than one interocular separation (Fig. 1d). The 2-s movement trajectory was windowed with a high-powered gaussian to smooth out the beginning and end of the movement. The axis of platform motion was chosen according to the visual direction tuning of each MT neuron, such that random-dot motion oscillated along the neuron's preferred-null axis. Thus, all platform movements were along an axis in the frontoparallel plane. Motion of the random-dot patch was consistent with that produced by a stationary surface. Thus, the boundaries of the random-dot patch moved relative to the receptive field (typically by less than 25% of the receptive field diameter) during each trial. The excursion of the patch relative to the receptive field was matched for near and far simulated depths, so that it could not generate a depth-sign preference.

Two opposite phases of platform motion were used (Fig. 1d, thick curves), such that retinal image motion began in the neuron's preferred direction on one-half of the trials, and began in the null direction on the remaining half. Throughout the movement, the monkey's only task was to fixate the world-fixed

target, and successful completion of the trial required that his gaze remained within an electronic fixation window. To allow the monkey an opportunity to make an initial catch-up saccade (if necessary) at the onset of pursuit, the fixation window was initially 3.0–4.0° square and was shrunk to 1.5–2.0° after 250 ms. Horizontal and vertical eye position was monitored with a scleral search coil and captured at a sampling rate of 250 Hz. For illustrative purposes, eye traces in Fig. 1d and Supplementary Fig. 2 were smoothed with a boxcar filter (position 50 ms wide; velocity 200 ms wide).

In the Retinal Motion (RM) condition, the monkey remained stationary and we replicated the visual image seen during the MP condition by translation of the OpenGL camera along the same trajectory that the monkey followed in the MP condition. As the OpenGL camera translated, it also rotated to maintain 'aim' at the fixation target. This simulates the smooth tracking eye movements of the animal and generates visual stimuli that match those in the MP condition. The monkey performed six to ten repetitions of each stimulus condition at each simulated depth, randomly interleaved.

Analysis. Single-unit data were analysed with custom software written in Matlab (Mathworks Inc.). For generation of peri-stimulus time histograms (Fig. 2a, b), firing rate was computed in 50-ms bins. To quantify selectivity for depth sign, we combined data across the two phases of platform motion and computed a mean firing rate across the total duration of each trial. Spikes were counted within a temporal window that began 80 ms after stimulus onset and ended 80 ms after stimulus offset (to compensate for response latency). For each neuron we computed a DSDI:

$$\mathrm{DSDI} = \frac{1}{4}\sum_{i=1}^{4}\frac{R_{\mathrm{far}(i)} - R_{\mathrm{near}(i)}}{\left|R_{\mathrm{far}(i)} - R_{\mathrm{near}(i)}\right| + \sigma_{\mathrm{av}}}$$

For each pair of depths symmetrical around zero (for example, ±1°), we calculated the difference in response between far (R_{far}) and near (R_{near}) relative to response variability (σ_{av}, the average standard deviation of the two responses). We then averaged across the four matched pairs of depths to obtain the DSDI, which ranges from −1 to 1. Neurons that respond more strongly to near stimuli will have negative values, whereas neurons that prefer far stimuli will have positive values. A DSDI was calculated separately for the MP and RM conditions. DSDIs were classified as significantly different from zero (or not) by permutation test (1,000 permutations, $P < 0.05$).

The above metric has the advantage of taking into account trial-to-trial variations in response while quantifying the magnitude of response differences between near and far. However, we also analysed our data by using a standard contrast measure and obtained very similar results (Supplementary Fig. 7).

Tuning curves in the MP condition were classified as monotonic if one of the following was true: the mean firing rates were a strictly monotonic function of equivalent disparity (that is, steadily increasing or decreasing), a second-order polynomial ($y = a_1 x + a_2 x^2 + a_3$) did not provide a significantly better fit than a linear function ($y = a_1 x + a_2$) by sequential F-test ($P > 0.05$), or the best-fitting second-order polynomial was itself strictly monotonic within the range of the data. Tuning curves not meeting any of these conditions were classified as non-monotonic (see, for example, Fig. 3a).

29. Gu, Y., DeAngelis, G. C. & Angelaki, D. E. A functional link between area MSTd and heading perception based on vestibular signals. *Nature Neurosci.* **10**, 1038–1047 (2007).

30. Nguyenkim, J. D. & DeAngelis, G. C. Disparity-based coding of three-dimensional surface orientation by macaque middle temporal neurons. *J. Neurosci.* **23**, 7117–7128 (2003).

LETTERS

Tumour maintenance is mediated by eNOS

Kian-Huat Lim[1]*†, Brooke B. Ancrile[1]*, David F. Kashatus[1]* & Christopher M. Counter[1]

Tumour cells become addicted to the expression of initiating oncogenes like *Ras*, such that loss of oncogene expression in established tumours leads to tumour regression[1]. HRas, NRas or KRas are mutated to remain in the active GTP-bound oncogenic state in many cancers[2]. Although Ras activates several proteins to initiate human tumour growth, only PI3K, through activation of protein kinase B (PKB; also known as AKT), must remain activated by oncogenic Ras to maintain this growth[3]. Here we show that blocking phosphorylation of the AKT substrate, endothelial nitric oxide synthase (eNOS or NOS3), inhibits tumour initiation and maintenance. Moreover, eNOS enhances the nitrosylation and activation of endogenous wild-type Ras proteins, which are required throughout tumorigenesis. We suggest that activation of the PI3K–AKT–eNOS–(wild-type) Ras pathway by oncogenic *Ras* in cancer cells is required to initiate and maintain tumour growth.

The reduction of *Ras* oncogene dependence to activation of AKT appears to be a consequence of redundant signalling provided by the established tumour microenvironment. Indeed, activation of AKT fosters tumorigenic growth of otherwise non-tumorigenic cells, provided such cells are mixed with tumour cells to establish the tumour microenvironment[3]. We exploited this cell-mixing assay to interrogate the signalling pathway downstream of AKT required for tumour maintenance. Although AKT can phosphorylate several proteins[4], we focused on BAD, FOXO, IKKα, TSC2 and eNOS, as the consequence of AKT phosphorylation of these proteins is not redundant with the functions of the oncoproteins expressed in cells used in the cell-mixing assay of tumour maintenance[5] (Supplementary Fig. 1). Non-tumorigenic PI3K-TtH^{LacZ} cells, derived from normal human kidney cells transformed by oncoproteins T/t-Ags and immortalized by hTERT (hereafter termed TtH cells), and which expressed p110-CAAX (to activate the PI3K-AKT pathway) and LacZ (to demark the cells in the tumour), had Bcl-X$_L$ short hairpin RNA (shRNA), eNOS shRNA or, as reported by others[6–8], dominant-acting FOXO3a-A3 or TSC2^{SA,TA} (mutated at AKT phosphorylation sites), or IKKα^{K44A} (kinase-inactive) proteins expressed to suppress the effects of AKT on these individual pathways (Supplementary Fig. 1). Knockdown of Bcl-X$_L$ and eNOS, ectopic expression and nuclear localization of FOXO3a-A3 (ref. 6), ectopic expression of TSC2^{SA,TA} leading to mTOR repression, as assessed by decreased S6k phosphorylation[7], and ectopic expression of IKKα^{K44A} leading to repression of NF-κB, as assessed by nuclear exclusion of p65[8], were validated (Supplementary Fig. 2). These five cell lines were mixed with tumorigenic HRas^{G12V}-transformed TtH cells (termed Ras^{G12V}-TtH cells) to establish a tumour microenvironment, injected into mice, and assayed for their contribution to the resultant tumour mass by treating tumours or derived tumour cells with X-gal to stain LacZ$^+$ cells blue. Positive control vector PI3K-TtH^{LacZ} cells extensively populated tumours, whereas negative control vector TtH^{LacZ} cells contributed little to the tumour mass, as evidenced by the prominent or weak blue staining, respectively. Expression of IKKα^{K44A} had little

effect; Bcl-X$_L$ shRNA, FOXO3a-A3 and TSC2^{SA,TA} had a mild effect; however, eNOS shRNA had the greatest effect on reducing the contribution of PI3K-TtH^{LacZ} (blue) cells in tumours (Fig. 1a).

To test if AKT phosphorylation, not just expression of eNOS, is required for tumour maintenance, AKT was validated to phosphorylate S$_{1177}$ of eNOS[9–11]. This was evidenced by a loss of S$_{1177}$ phosphorylation of endogenous eNOS upon pharmacological inhibition of AKT signalling with LY294002, or by mutating S$_{1177}$ of ectopic eNOS to alanine in PI3K-TtH^{LacZ} cells (Fig. 1b). PI3K-TtH^{LacZ} cells in which eNOS was knocked down were then engineered to express eNOS^R in the wild-type or S$_{1177}$A mutant configuration (Supplementary Fig. 3), and assayed for tumour maintenance by the aforementioned cell-mixing assay. Control PI3K-TtH^{LacZ} cells populated tumours, and this contribution was greatly reduced upon knockdown of eNOS, as evidenced by the reduction in blue staining. This loss was rescued by wild-type, but not S$_{1177}$A, mutant eNOS^R (Fig. 1c). Thus, activation of the PI3K–AKT–eNOS pathway promotes tumour maintenance.

eNOS has been detected in tumour cells[12], and catalyses the synthesis of nitric oxide. This can facilitate *S*-nitrosylation of the thiol group of cysteines in proteins[13], such as that of C$_{118}$ of HRas, which enhances the dissociation of guanine nucleotides, thereby increasing GTP-bound HRas[14]. Wild-type Ras proteins can be required for activation of the MAPK pathway by oncogenic Ras[15] and membrane targeting of RasGAP, which inhibits wild-type but not oncogenic Ras, reverts oncogenic Ras transformation of NIH3T3 cells[16]. This suggests that wild-type Ras proteins may facilitate oncogenic signalling. Collectively, we speculated that AKT activation of eNOS maintains tumour growth in the absence of oncogenic Ras by activating wild-type Ras through *S*-nitrosylation of C$_{118}$. To test this, activated AKT in PI3K-TtH was shown to foster HRas nitrosylation through eNOS. Specifically, most HRas nitrosylation was lost by treatment with the PI3K inhibitor wortmannin (Fig. 2a), by mutating C$_{118}$ in HRas to serine (a minor change that exchanges the sulphur atom for oxygen but nevertheless blocks nitrosylation[14]) (Fig. 2b), or by knocking down eNOS (Fig. 2c). Conversely, HRas nitrosylation was elevated upon activation of AKT by p110-CAAX (Fig. 2c). Reduction of HRas nitrosylation by eNOS shRNA also reduced levels of active GTP-bound HRas (Fig. 2c). Because TtH cells express HRas and NRas, but not KRas (not shown), and C$_{118}$ is conserved among all Ras proteins, we tested and confirmed that activated AKT in PI3K-TtH cells also led to elevated levels of nitrosylated and GTP-bound endogenous NRas, which were reduced upon knockdown of eNOS (Fig. 2c). Thus, AKT activation of eNOS promotes nitrosylation and activation of wild-type Ras proteins. Then, to assess the biological consequence of *S*-nitrosylation of wild-type HRas in tumour maintenance, we tested whether replacing endogenous wild-type HRas with the nitrosylation-resistant C$_{118}$S mutant version reduced tumour maintenance. HRas was knocked down by shRNA in PI3K-TtH^{LacZ} cells and complemented with vector encoding an shRNA-resistant
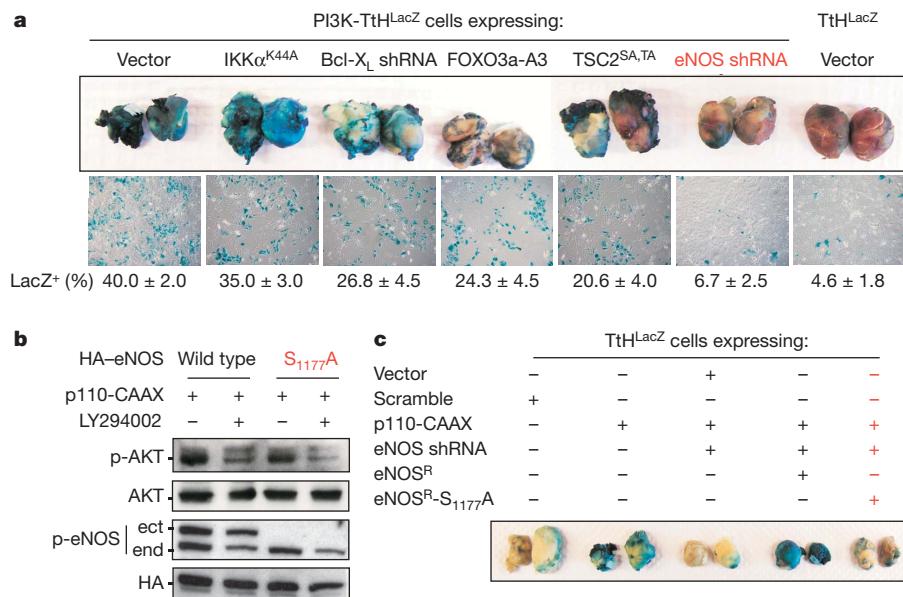
Figure 1 | AKT promotes tumour maintenance by phosphorylation of eNOS. **a**, PI3K-TtH[LacZ] or **c**, TtH[LacZ] cells expressing indicated constructs, were mixed with Ras[G12V]-TtH cells, injected into mice, and tumours or re-cultured tumour cells were stained with X-gal ($n = 5$, mean ± s.e.m.).

b, Protein levels of phosphorylated AKT (p-AKT), phosphorylated eNOS (p-eNOS; ect, ectopic; end, endogenous), HA–eNOS (HA) and AKT in PI3K-TtH[LacZ] cells expressing wild-type or S$_{1177}$A HA–eNOS treated with DMSO or LY294002.

HRas (HRas[R]) in the wild-type or C$_{118}$S mutant configuration that reduced GTP loading, or, as a control, no transgene (Supplementary Fig. 4). These cell lines, or as control cells expressing either a scramble control sequence or HRas shRNA alone, were mixed with Ras[G12V]-TtH cells, injected into mice, and the resultant tumours assayed for the presence of blue LacZ[+] cells as a measure of tumour maintenance. Knockdown of wild-type HRas reduced the ability of PI3K-AKT signalling to foster tumour maintenance, as evidenced by a sixfold reduction of PI3K-TtH[LacZ] (blue) cells in the tumours. This effect was reversed upon expressing wild-type HRas[R], but less so with the C$_{118}$S mutant version of HRas[R] (Fig. 2d). Thus, activation of the PI3K-AKT-eNOS pathway promotes tumour maintenance by S-nitrosylation and activation of wild-type Ras.

As oncogenic Ras must activate the PI3K-AKT pathway both to initiate and maintain tumour growth[3], we tested whether AKT-mediated activation of eNOS was also required for the establishment

of tumours. A scramble control or eNOS shRNA was introduced into tumorigenic Ras[G12V]-TtH cells, and knockdown of eNOS complemented by RNAi-resistant eNOS (eNOS[R]) in the wild-type or S$_{1177}$A mutant configuration resistant to S$_{1177}$ phosphorylation (Supplementary Fig. 5). These four cell lines were injected into mice, and tumour growth monitored over time. Scramble control cells rapidly formed tumours, whereas tumour growth was almost abolished upon knockdown of eNOS. This loss of tumour growth was rescued by the wild-type, but not the S$_{1177}$A version, of eNOS[R] (Fig. 3a and Supplementary Fig. 6), indicating that S$_{1177}$ phosphorylation of eNOS is required for tumour initiation and maintenance. These results were validated in a chemical carcinogen-induced spontaneous Ras-driven cancer model. DMBA (7,12-dimethylbenz(a)anthracene) followed by 12-O-tetradecanoylphorbol-13-acetate (TPA) were topically applied to eNOS[+/+] and eNOS[−/−] mice to induce skin papillomas characterized by Ras oncogenic mutations[17]. The result was an
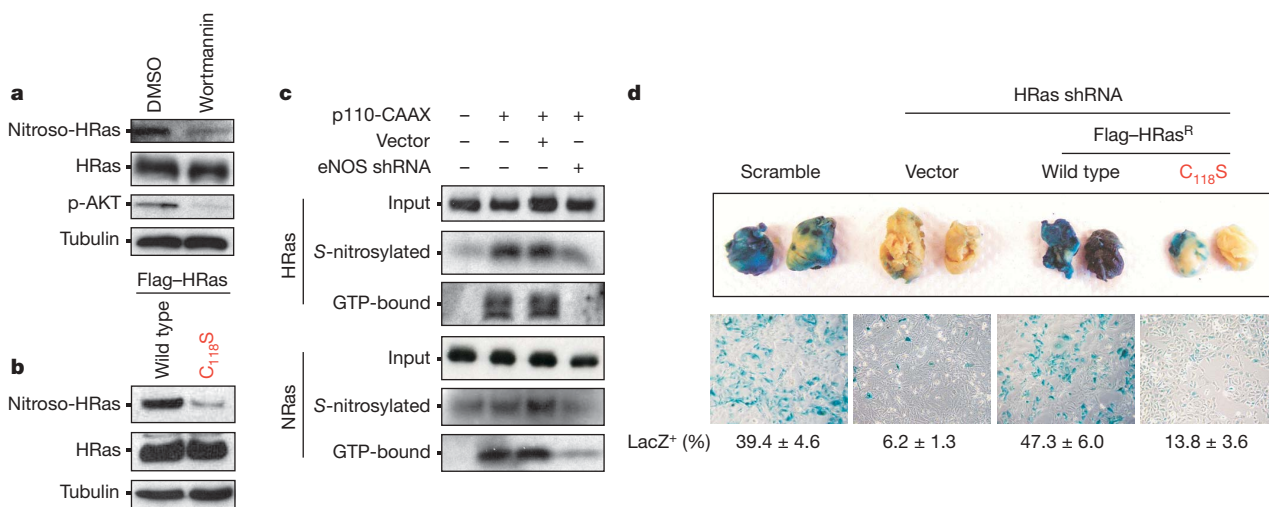

Figure 2 | eNOS activates wild-type HRas to promote tumour maintenance. **a–c**, Protein levels of S-nitrosylated (nitroso), GTP-bound, total or input HRas or NRas, phosphorylated AKT (p-AKT), or as a loading control tubulin in PI3K-TtH[LacZ] cells treated with DMSO or wortmannin (**a**), PI3K-TtH[LacZ] cells transfected with wild-type or C$_{118}$S HRas (**b**), or

TtH[LacZ] cells expressing the indicated combinations of p110-CAAX, vector or eNOS shRNA (**c**). **d**, Ras[G12V]-TtH cells were mixed with PI3K-TtH[LacZ] cells expressing the indicated constructs, injected into mice, and tumours or re-cultured tumour cells were stained with X-gal to visualize PI3K-TtH[LacZ] cells $n = 5$, mean ± s.e.m.).

approximate threefold drop in the number of tumours per $eNOS^{-/-}$ mouse (Fig. 3b and Supplementary Fig. 7). Thus, independent models of cancer demonstrate eNOS is required for tumorigenesis.

To test whether eNOS mediates oncogenic Ras signalling in a cancer associated with the most commonly mutated Ras family member, $KRas^2$, the amount of activated (S_{1177} phosphorylated) eNOS was first assayed in cancer cell lines and tumour specimens isolated from patients diagnosed with pancreatic cancer. Compared with normal tissue specimens, CFPac-1, MIAPaCa-2 and Capan-1 cells exhibited the highest level of S_{1177} phosphorylation of eNOS (Fig. 4a). Activated KRas and S_{1177}-phosphorylated eNOS were also elevated in the tumour specimens compared with matched and unmatched normal tissue controls (Fig. 4b and ref. 18), with the caveat that biopsies also contained stromal tissue that could have contributed to detected eNOS phosphorylation.

Next, we tested whether eNOS was required for pancreatic tumour growth. Knockdown of eNOS reduced tumour growth by 50-fold in CFPac-1 cells, and tumour size of MIAPaCa-2 cells by 80% (Fig. 4c, and Supplementary Figs 8 and 9). As a control, knockdown of eNOS in cell lines AsPc-1 and SW1990, which exhibited poor eNOS activation, had no obvious effect on tumour growth (data not shown). Using a more direct assay for tumour maintenance, *in vivo* doxycycline (dox)-induced shRNA knockdown of eNOS in CFPac-1 and MIAPaCa-2 cells after tumours were established inhibited tumour growth, as evidenced by reduced size and/or gross necrosis of tumours excised at the termination of the experiment (Fig. 4d). In some mice, this eventually led to tumour regression (data not shown). Thus, eNOS is required both to initiate and maintain tumour growth of these human pancreatic cancer cells.

To test whether eNOS promoted tumour growth through nitrosylation of Ras in pancreatic cancer cells, we determined which Ras family members were inactivated by eNOS shRNA in CFPac-1 and MIAPaCa-2 cells. Not surprisingly, GTP-bound KRas was unchanged on knockdown of eNOS (Fig. 4c and Supplementary Fig. 9), as *KRas* is
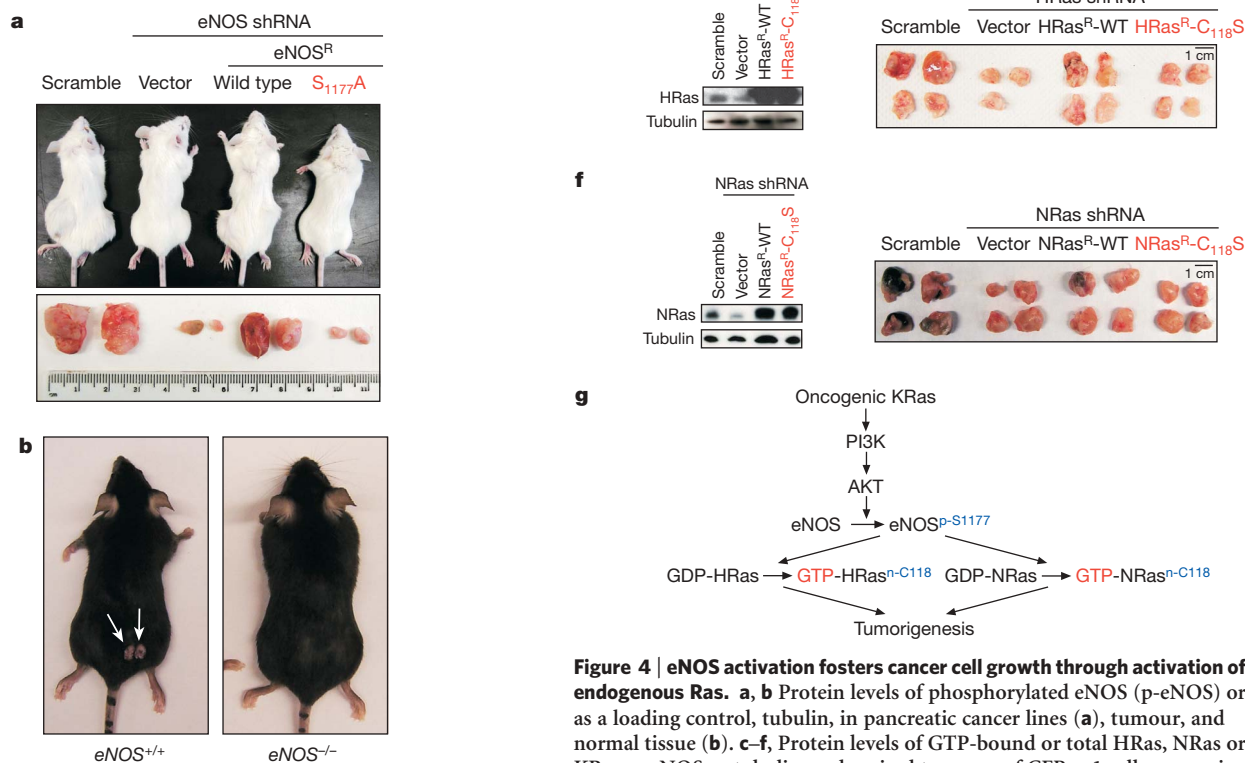

Figure 3 | eNOS activation is required for tumour growth. Representative mice and/or tumours: **a**, after injection with Ras^{G12V}-TtH cells expressing indicated constructs; or **b**, after treatment with DMBA/TPA to induce skin tumours on mice of indicated genotype (week 20).

Figure 4 | eNOS activation fosters cancer cell growth through activation of endogenous Ras. a, b Protein levels of phosphorylated eNOS (p-eNOS) or as a loading control, tubulin, in pancreatic cancer lines (**a**), tumour, and normal tissue (**b**). **c–f**, Protein levels of GTP-bound or total HRas, NRas or KRas, p-eNOS or tubulin, and excised tumours of CFPac-1 cells expressing eNOS or scramble shRNA (**c**), CFPac-1 and MIAPaCa-2 cells expressing dox-inducible shRNA +/− dox (**d**), and CFPac-1 cells expressing HRas (**e**) or NRas shRNA (**f**) plus a vector or an RNAi-resistant wild-type or C_{118}S HRas or NRas, or a scramble sequence. **g**, Proposed signalling.

mutated to remain active in these two cell lines[19,20]. Consistent with this, oncogenic KRas harbouring the $C_{118}S$ mutation remained tumorigenic (Supplementary Fig. 10), pointing towards wild-type Ras proteins as the target of eNOS signalling. Indeed, GTP-bound endogenous wild-type HRas and NRas were reduced upon shRNA knockdown of eNOS. Moreover, as the wild-type allele of *KRas* is deleted in MIAPaCa-2 cells[20], several oncogenic *KRas*-positive cell lines[20] and tumour tissues[21], wild-type HRas and NRas, but not KRas, appear to be the targets of eNOS signalling in pancreatic cancer cells (Fig. 4c and Supplementary Fig. 9).

To test if activation of HRas or NRas by eNOS is required for pancreatic tumour growth, HRas or NRas were knocked down by shRNA in CFPac-1 and/or MIAPaCa-2 cells and complemented with an HRas or NRas that was engineered to be resistant to RNAi (HRas^R, NRas^R) in the wild-type or $C_{118}S$ mutant configuration. Resultant cells were then assayed for tumour growth in mice. Positive control, scramble-treated CFPac-1 and/or MIAPaCa-2 cells readily formed tumours in mice, whereas this growth was reduced when endogenous HRas, and to a lesser degree NRas, was knocked down. This loss of tumour growth was rescued by expressing the appropriate wild-type HRas or NRas, but not the $C_{118}S$ nitrosylation mutants (Fig. 4e, f and Supplementary Figs 11–13). Similar results were found when the cells were assayed for transformed growth *in vitro*, suggestive of a tumour-cell autonomous defect when wild-type Ras proteins cannot be nitrosylated (Supplementary Fig. 14). Thus, oncogenic KRas-driven pancreatic cancer tumour growth was mediated by eNOS nitrosylation of endogenous wild-type HRas and NRas (Fig. 4g).

In summary, we demonstrate that the continual need for PI3K-AKT signalling during initiation and maintenance of oncogenic Ras-driven tumour growth is due, at least in part, to activation of eNOS through phosphorylation of S_{1177}. This in turn leads to *S*-nitrosylation at C_{118} and correspondingly activation of the other wild-type Ras family members, perhaps as a means to diversify the Ras signal beyond that of oncogenic Ras (Fig. 4g). In agreement, the wild-type counterpart of oncogenic Ras is not required for tumorigenesis (Supplementary Fig. 15), and is even deleted in some tumours[21,22]; whereas wild-type HRas and NRas are required for oncogenic KRas-driven tumour growth, and appear to have non-redundant activities[23–26]. Effects of eNOS on tumorigenesis have been largely attributed to its activity in endothelial cells[12]. Our results now suggest a key role for tumour-expressed eNOS in the tumorigenic process. Because eNOS plays multiple roles in tumorigenesis[12], and delivery of a peptide fragment of the protein cavtratin, which can inhibit eNOS, displays anti-tumour activity[27], we speculate that inhibition of eNOS, perhaps in combination with inhibition of wild-type Ras protein function or processing[2], could have therapeutic value in the treatment of oncogenic Ras-driven human cancers such as those of the pancreas.

METHODS SUMMARY

TtH and the pancreatic cancer cell lines were stably infected retroviruses encoding the indicated shRNAs, transgenes or no insert as described[5], and appropriate expression verified by immunoblot or polymerase chain reaction with reverse transcription (RT–PCR). Detection of GTP-bound or nitrosylated Ras was performed as described in Methods. One or a mixture of two cell lines were injected into the flanks of immunocompromised mice to assay for tumour growth; where indicated, they were excised and assayed for LacZ-positive cells, as described[3]. Induction of shRNA *in vivo* by dox[3] and DMBA/TPA treatments[28] were performed as described.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Giuriato, S. *et al.* Conditional animal models: a strategy to define when oncogenes will be effective targets to treat cancer. *Semin. Cancer Biol.* 14, 3–11 (2004).

2. Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nature Rev. Cancer* 3, 11–22 (2003).

3. Lim, K. H. & Counter, C. M. Reduction in the requirement of oncogenic Ras signaling to activation of PI3K/AKT pathway during tumor maintenance. *Cancer Cell* 8, 381–392 (2005).

4. Luo, J., Manning, B. D. & Cantley, L. C. Targeting the PI3K-Akt pathway in human cancer: rationale and promise. *Cancer Cell* 4, 257–262 (2003).

5. O'Hayer, K. M. & Counter, C. M. A genetically defined normal somatic human cell system to study ras oncogenesis *in vitro* and *in vivo*. *Methods Enzymol.* 407, 637–647 (2006).

6. Brunet, A. *et al.* Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. *Cell* 96, 857–868 (1999).

7. Manning, B. D. *et al.* Identification of the tuberous sclerosis complex-2 tumor suppressor gene product tuberin as a target of the phosphoinositide 3-kinase/akt pathway. *Mol. Cell* 10, 151–162 (2002).

8. Regnier, C. H. *et al.* Identification and characterization of an IκB kinase. *Cell* 90, 373–383 (1997).

9. Michell, B. J. *et al.* The Akt kinase signals directly to endothelial nitric oxide synthase. *Curr. Biol.* 9, 845–848 (1999).

10. Fulton, D. *et al.* Regulation of endothelium-derived nitric oxide production by the protein kinase Akt. *Nature* 399, 597–601 (1999).

11. Dimmeler, S. *et al.* Activation of nitric oxide synthase in endothelial cells by Akt-dependent phosphorylation. *Nature* 399, 601–605 (1999).

12. Fukumura, D., Kashiwagi, S. & Jain, R. K. The role of nitric oxide in tumour progression. *Nature Rev. Cancer* 6, 521–534 (2006).

13. Hess, D. T. *et al.* Protein S-nitrosylation: purview and parameters. *Nature Rev. Mol. Cell Biol.* 6, 150–166 (2005).

14. Lander, H. M. *et al.* Redox regulation of cell signalling. *Nature* 381, 380–381 (1996).

15. Hamilton, M. & Wolfman, A. Ha-ras and N-ras regulate MAPK activity by distinct mechanisms *in vivo*. *Oncogene* 16, 1417–1428 (1998).

16. Huang, D. C., Marshall, C. J. & Hancock, J. F. Plasma membrane-targeted ras GTPase-activating protein is a potent suppressor of p21ras function. *Mol. Cell. Biol.* 13, 2420–2431 (1993).

17. Quintanilla, M. *et al.* Carcinogen-specific mutation and amplification of Ha-ras during mouse skin carcinogenesis. *Nature* 322, 78–80 (1986).

18. Lim, K. H. *et al.* Divergent roles for RalA and RalB in malignant growth of human pancreatic carcinoma cells. *Curr. Biol.* 16, 2385–2394 (2006).

19. Moore, P. S. *et al.* Genetic profile of 22 pancreatic carcinoma cell lines. Analysis of K-ras, p53, p16 and DPC4/Smad4. *Virchows Arch.* 439, 798–802 (2001).

20. Kita, K. *et al.* Growth inhibition of human pancreatic cancer cell lines by anti-sense oligonucleotides specific to mutated K-ras genes. *Int. J. Cancer* 80, 553–558 (1999).

21. Wan, J. *et al.* Loss of heterozygosity of Kras2 gene on 12p12–13 in Chinese colon carcinoma patients. *World J. Gastroenterol.* 12, 1033–1037 (2006).

22. Li, J. *et al.* LOH of chromosome 12p correlates with Kras2 mutation in non-small cell lung cancer. *Oncogene* 22, 1243–1246 (2003).

23. Parikh, C., Subrahmanyam, R. & Ren, R. Oncogenic NRAS, KRAS, and HRAS exhibit different leukemogenic potentials in mice. *Cancer Res.* 67, 7139–7146 (2007).

24. Esteban, L. M. *et al.* Targeted genomic disruption of H-ras and N-ras, individually or in combination, reveals the dispensability of both loci for mouse growth and development. *Mol. Cell. Biol.* 21, 1444–1452 (2001).

25. Johnson, L. *et al.* K-ras is an essential gene in the mouse with partial functional overlap with N-ras. *Genes Dev.* 11, 2468–2481 (1997).

26. Fotiadou, P. P. *et al.* Wild-Type NRas and KRas perform distinct functions during transformation. *Mol. Cell. Biol.* 27, 6742–6755 (2007).

27. Gratton, J. P. *et al.* Selective inhibition of tumor microvascular permeability by cavtratin blocks tumor progression in mice. *Cancer Cell* 4, 31–39 (2003).

28. Ancrile, B., Lim, K. H. & Counter, C. M. Oncogenic Ras-induced secretion of IL6 is required for tumorigenesis. *Genes Dev.* 21, 1714–1719 (2007).

Author Contributions K.-H.L., B.B.A. and D.F.K. performed the experiments. All authors conceived and designed experiments and participated in the writing of the paper.

nature

METHODS

Plasmids. pBabepuro, neo, bleo and hygro were used as control vectors[3]. The following complementary DNAs (cDNAs) were subcloned into one of the aforementioned pBabe vectors: HA–IKKα[K44A] cDNA[29], FOXO3a-A3 cDNA[30] engineered with an amino (N)-terminal HA tag, Flag–TSC2[S393A,T1462A] cDNA[7] (termed here as TSC2[SA,TA]), eNOS cDNA engineered with a carboxy (C)-terminal HA tag and to be resistant to shRNA by introducing the three silent mutations $G_{1821} \rightarrow A$, $T_{1827} \rightarrow C$ and $G_{1830} \rightarrow A$ alone (eNOS[R]) or in conjunction with the mutation $A_{3519}GC \rightarrow GCC$ that altered S_{1177} to A ($S_{1177}A$ eNOS[R]), and wild-type Flag-epitope tagged HRas or NRas cDNAs engineered to be resistant to shRNA by introducing the silent mutations in the region targeted by RNAi (Flag–HRas[R]; Flag–NRas[R]) alone or in conjunction with the mutation $T_{342}GT \rightarrow TCT$ ($C_{118}S$ Flag–HRas[R], $C_{118}S$ Flag–NRas[R]) that altered C_{118} to S. Bcl-X$_L$ shRNA (5′-AGCGTAGACAAGGAGATGC), eNOS shRNA (5′-AAGAGTTATAAGATCCGCTTC), HRas shRNA (5′-GGCAAGAGTG-CGCTGACCATC), NRas shRNA (5′-CAAGAAGAGTACAGTGCCATG) or eNOS scramble control (5′-AAGCGTTAAAAGATCCGCTTC) sequences were cloned into pSUPER-PURO-RETRO (Oligoengine). The plasmid system for dox-inducible shRNA[3] was adapted to encode eNOS shRNA.

Cell lines. TtH and the pancreatic cancer cell lines were previously described[31]. Derived lines were generated by stable infection with the indicated combinations of amphotrophic retroviruses generated from the aforementioned pBabe plasmids, as previously described[5].

Cell treatments. Cells were treated with LY294002 (Cell Signaling Technologies) or wortmannin (Sigma) at a final concentration of 20 μM or 10 nM, respectively, for 1 h before analysis.

Immunoblotting. HA–IKKα[K44A], FOXO3a-A3–HA, HA–eNOS or variants thereof, Flag–TSC2[SA,TA], endogenous Bcl-X$_L$, p70 S6 kinase, T_{389} phosphorylated p70 S6 kinase, HRas, KRas or NRas, S_{1177} phosphorylated eNOS (both to detect activated eNOS and assess eNOS expression), S_{473} phosphorylated AKT, actin, p65 and tubulin were detected by immunoblotting with anti-HA (Roche), anti-Flag (Sigma), anti-Bcl-xL, anti-p70 S6 Kinase, anti-Thr389 Phospho-p70 S6 Kinase, anti-Ser1177 Phospho-eNOS, anti-Ser473 Phospho-AKT (Cell Signaling Technology), anti-HRas, anti-KRas, anti-NRas, anti-actin (Santa Cruz), anti-p65 (Rockland) and anti-tubulin (Sigma) antibodies, respectively.

RT–PCR. eNOS and GAPDH mRNA was RT–PCR amplified with the primers 5′-CAGTGTCCAACATGCTGCTGGAAATTG and 5′-TAAAGGTCTTCTT-CCTGGTGATGC, and the primers 5′-ACCACAGTCCATGCCATCAC and 5′-TCCACCACCCTGTTGCTGTA, respectively.

GTP and nitrosylated Ras. GTP-bound or nitrosylated Ras were captured as previously described[32,33] and immunoblotted with either an anti-Flag (Sigma) or an anti-HRas, anti-KRas or anti-NRas (Santa Cruz) antibody to detect Flag–H-Ras or endogenous H, N or KRas proteins, respectively.

Soft agar. Soft agar assays were done in triplicate and twice independently as previously described[5].

Tumour growth. As previously described[3], the tested cell line (tumour initiation) or a mixture of two cell lines (cell mixing assay for tumour maintenance) were injected subcutaneously into four flanks of SCID/Beige mice. For tumour initiation experiments, tumours were removed and photographed when control tumours reached maximum volume. For cell mixing assays, the four tumours were removed when they reached maximum volume, human cells derived from the two tumours by re-culture in selective media (G418), and the two other whole tumours were treated with X-gal to stain LacZ[+] cells blue and photographed. CFPac-1 and MIAPaCa-2 cells engineered to contain a dox-inducible eNOS shRNA[3] were injected into both flanks of five SCID/Beige mice. Tumours were allowed to reach a diameter of 0.6 cm, after which three mice were provided with doxycycline in their diet and two mice left untreated for 11 days (CFPac-1 cells) or 13 days (MIAPaCa-2 cells), after which tumours were removed and photographed. DMBA/TPA treatments were performed as previously described[28] on 15 eNOS[+/+] C57BL/6J and 15 eNOS[−/−] C57BL/6J (B6.129P2-*Nos3*[tm1Unc]/J) mice[34] (Jackson Laboratory). All animal work was approved by the Duke University Medical Center Institutional Animal Care and Use Committee.

Tumour and normal human specimens. Flash-frozen tissue samples were provided devoid of all identifying information under a Duke University Medical Center approved Institutional Review Board protocol.

29. Woronicz, J. D. *et al.* IκB kinase-β: NF-κB activation and complex formation with IκB kinase-α and NIK. *Science* **278**, 866–869 (1997).

30. Hu, M. C. *et al.* IκB kinase promotes tumorigenesis through inhibition of forkhead FOXO3a. *Cell* **117**, 225–237 (2004).

31. Lim, K. H. *et al.* Activation of RalA is critical for Ras-induced tumorigenesis of human cells. *Cancer Cell* **7**, 533–545 (2005).

32. de Rooij, J. & Bos, J. L. Minimal Ras-binding domain of Raf1 can be used as an activation-specific probe for Ras. *Oncogene* **14**, 623–625 (1997).

33. Jaffrey, S. R. *et al.* Protein S-nitrosylation: a physiological signal for neuronal nitric oxide. *Nature Cell Biol.* **3**, 193–197 (2001).

34. Shesely, E. G. *et al.* Elevated blood pressures in mice lacking endothelial nitric oxide synthase. *Proc. Natl Acad. Sci. USA* **93**, 13176–13181 (1996).

LETTERS

Cutaneous cancer stem cell maintenance is dependent on β-catenin signalling

Ilaria Malanchi[1], Hector Peinado[2], Deepika Kassen[1], Thomas Hussenet[1], Daniel Metzger[3], Pierre Chambon[3], Marcel Huber[4], Daniel Hohl[4], Amparo Cano[2], Walter Birchmeier[5] & Joerg Huelsken[1]

Continuous turnover of epithelia is ensured by the extensive self-renewal capacity of tissue-specific stem cells[1]. Similarly, epithelial tumour maintenance relies on cancer stem cells (CSCs), which co-opt stem cell properties[2]. For most tumours, the cellular origin of these CSCs and regulatory pathways essential for sustaining stemness have not been identified. In murine skin, follicular morphogenesis is driven by bulge stem cells that specifically express CD34. Here we identify a population of cells in early epidermal tumours characterized by phenotypic and functional similarities to normal bulge skin stem cells. This population contains CSCs, which are the only cells with tumour initiation properties. Transplants derived from these CSCs preserve the hierarchical organization of the primary tumour. We describe β-catenin signalling[3] as being essential in sustaining the CSC phenotype. Ablation of the β-catenin gene results in the loss of CSCs and complete tumour regression. In addition, we provide evidence for the involvement of increased β-catenin signalling in malignant human squamous cell carcinomas. Because Wnt/β-catenin signalling is not essential for normal epidermal homeostasis, such a mechanistic difference may thus be targeted to eliminate CSCs[4] and consequently eradicate squamous cell carcinomas.

Two populations of epithelial stem cells have been identified in murine skin[1,5]. One is located in the bulge region of hair follicles and is responsible for follicular homeostasis. These bulge stem cells can be isolated by fluorescence-activated cell sorting (FACS) based on CD34 expression[6,7] with concurrent exclusion of dead or non-keratinocyte cells by using 7-amino-actinomycin D (7-AAD), CD31 and CD45 (Fig. 1a, e). Alternatively, and with identical results, we isolate bulge stem cells as CD34$^+$ yellow-fluorescent-protein-positive (YFP$^+$) cells from *K14-cre:ROSA26-lox-stop-lox-YFP* mice, which express YFP in keratinocyte lineages only (data not shown). CD34-positive bulge stem cells account for about $1.8 \pm 0.2\%$ ($n = 4$; mean ± s.d.) of keratinocytes in normal skin. In cutaneous tumours derived by chemical (dimethylbenz[a]anthracene/12-O-tetradecanoylphorbol-13-acetate (DMBA/TPA)) carcinogenesis we observe a nine-fold increase in the CD34$^+$ or CD34$^+$YFP$^+$ population (Fig. 1b, d; $n = 4$; $17.2 \pm 4.8\%$ (mean ± s.d.) of keratinocyte-lineage-derived tumour cells). CD34-expressing tumour cells are located in close contact to stromal areas (Fig. 1g, h) and also express additional, established markers of bulge skin stem cells[8,9] such as Gas6, ephrin A4, Sox9, Runx1, T-cell-specific transcription factor 3 (TCF-3) and tenascin C (Fig. 1i–m and Supplementary Fig. 1). Although the DMBA/TPA model generates mainly benign papillomatous lesions[10], occasionally progression to squamous cell carcinomas (SCCs) is observed. Within these advanced tumours the CD34$^+$

population is maintained (Fig. 1d, g) and expression of CD34 is often found at the invading front, in line with recent reports identifying CSCs as the main source of tumour invasion[11,12]. CD34 seems also to be of functional importance in skin tumour formation[6], although its mode of action is currently not understood. Together these data identify a population of cells in early and advanced epidermal skin tumours that closely resembles that of bulge stem cells.

The *in vivo* tumorigenic capacity of CD34$^+$ cells isolated by magnetic cell sorting (MACS) (Fig. 1c and Supplementary Fig. 2) was compared with that of unsorted tumour cells by serial dilution experiments in an orthotopic transplantation system, using tumour cells in conjunction with primary keratinocytes and dermal fibroblasts to permit whole-skin reconstitution. The CD34$^+$ population is over 100-fold more potent in initiating secondary tumours than unsorted tumour cells (Fig. 2a; $n = 50$). Secondary tumours maintain a stable population of CD34$^+$ cells that retain tumour initiation potential, giving rise to tertiary tumours (Fig. 2b and data not shown). In contrast, CD34$^-$ cells (Supplementary Fig. 2) never produce tumours ($n = 9$; 2.5×10^4 cells per transplant), although extensive proliferation is detected in this population, which presumably contains transit-amplifying cells (Fig. 2c–e). Secondary tumours derived from CD34$^+$ cells closely resemble the architecture of the parental tumour and maintain a small population of CD34$^+$ cells among the majority of keratin-10-expressing cells committed to differentiation (Fig. 2f, g). This ability to maintain their own population while producing differentiated progeny further substantiates the stem-cell nature of cells contained in the CD34$^+$ population. The relative abundance of this CSC-containing population and the enrichment of tumour-initiating cells is comparable to that of other solid tumours[13,14]. However, in contrast with these other systems, orthotopic transplantations in the skin are accompanied by an initial lack of nutrient supply and a strong inflammatory response—mainly by innate immune cells—which may explain why more tumour cells (unsorted or CD34$^+$) are required to generate secondary tumours in the skin. Our data, together with indirect evidence[15], strongly suggest that skin stem cells are potential targets of tumour initiation. We now provide compelling evidence for this idea by combining chemical carcinogenesis with lineage tracing of bulge stem cells, using *K15-creERT:ROSA26-lox-stop-lox-lacZ* mice as a genetic tool (Supplementary Fig. 3). Taken together, these data strongly suggest that bulge-derived CD34$^+$ stem cells are a source of tumorigenesis.

The Wnt/β-catenin signalling pathway controls the differentiation of skin stem cells without affecting epidermal homeostasis[16–19], and it has been implicated in a variety of human tumours[3,20]. We therefore examined the involvement of this pathway in chemical-derived skin

[1]École Polytechnique Fédérale de Lausanne (EPFL)/ISREC (Swiss Institute for Experimental Cancer Research) and National Center of Competence in Research (NCCR) 'Molecular Oncology', Chemin des Boveresses 155, 1066 Epalinges, Switzerland. [2]Departamento de Bioquímica, Instituto de Investigaciones Biomédicas 'Alberto Sols', CSIC-UAM, Arturo Duperier 4, 28029 Madrid, Spain. [3]Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), BP 10142, CU de Strasbourg, 67404 Illkirch, France. [4]Laboratory of Cutaneous Biology, Dermatology CHUV and FBM UNIL, Avenue de Beaumont 29, 1011 Lausanne, Switzerland. [5]Max Delbrück Centrum, Robert-Roessle-Strasse 10, 13122 Berlin, Germany.

tumours and assessed β-catenin signalling activity *in vivo* by using a reporter mouse strain with *lacZ* knock-in at the *conductin/axin2* locus, which is a well-established general target gene of Wnt signalling[21]. In contrast with the situation in normal epidermis (Fig. 3a), we observe strong signalling activity in tumours (Fig. 3b), in particular in the region where CD34[+] cells localize. Similarly, nuclear β-catenin, a hallmark of active signalling, is preferentially found in CD34[+] CSCs (Fig. 3c; 21.7 ± 8.4% (mean ± s.d.) of CD34[+] cells versus 3.3% ± 1.2% (mean ± s.d.) of CD34[−] cells; see ref. 12). This signalling activity prompted us to perform a dedicated tumour regression experiment by inducing *β-catenin* gene ablation in established

DMBA/TPA-derived tumours (*K14-creER^{T2}:β-catenin^{lox/lox}*; Fig. 3g). β-Catenin deletion results in complete regression, whereas the growth of control tumours is unaffected by the activation of creER[T2]. Regression is characterized by extensive terminal differentiation and the complete absence of mitotic figures within five to six weeks (Fig. 3d). Similarly, spontaneous regression was observed in a skin tumour model (TG.AC)[22] expressing activated *H-ras* (data not shown). In addition, deletion of β-catenin in the skin also prevents tumour initiation by chemical carcinogenesis or *H-ras* (Supplementary Figs 4 and 5). The first phenotype we detect in β-catenin-deficient tumours two weeks after deletion is a marked loss of the CD34[+] population of CSCs as analysed by FACS and bromodeoxyuridine (BrdU) long-term label-retaining assays (Fig. 3e, f; *n* = 7;
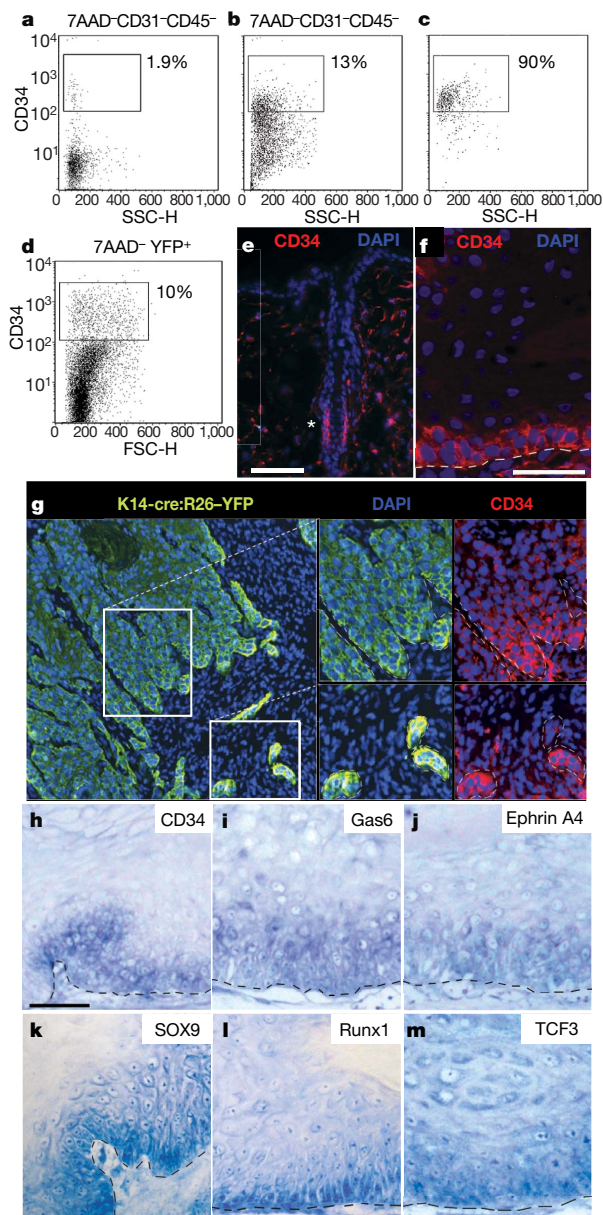
Figure 1 | Skin cancers contain cells with bulge stem cell phenotype.
a–d, Density plots of cells with stem cell phenotype defined as CD34[+] and viable (7AAD[−]) as well as CD31[−]CD45[−] or K14-cre:R26–YFP[+] (*n* = 4 each): adult mouse skin (**a**), murine papilloma (**b**) and murine SCC (**d**). **c**, Purity of CD34[+] tumour cells after enrichment by MACS. **e, f,** Immunofluorescence of CD34 in normal murine skin (**e**) and tumour (**f**). DAPI, 4,6-diamidino-2-phenylindole. **g,** CD34[+] cells (red) in murine SCCs as detected by immunofluorescence in K14-cre:R26YFP[+] mice. Insets: magnification with small groups of invading tumour cells. **h–m,** expression of bulge stem cell markers in papillomas as detected by RNA *in situ* hybridization: CD34 (**h**), Gas6 (**i**), ephrin A4 (**j**), SOX9 (**k**), Runx (**l**) and TCF-3 (**m**). Scale bars, 50 μm.
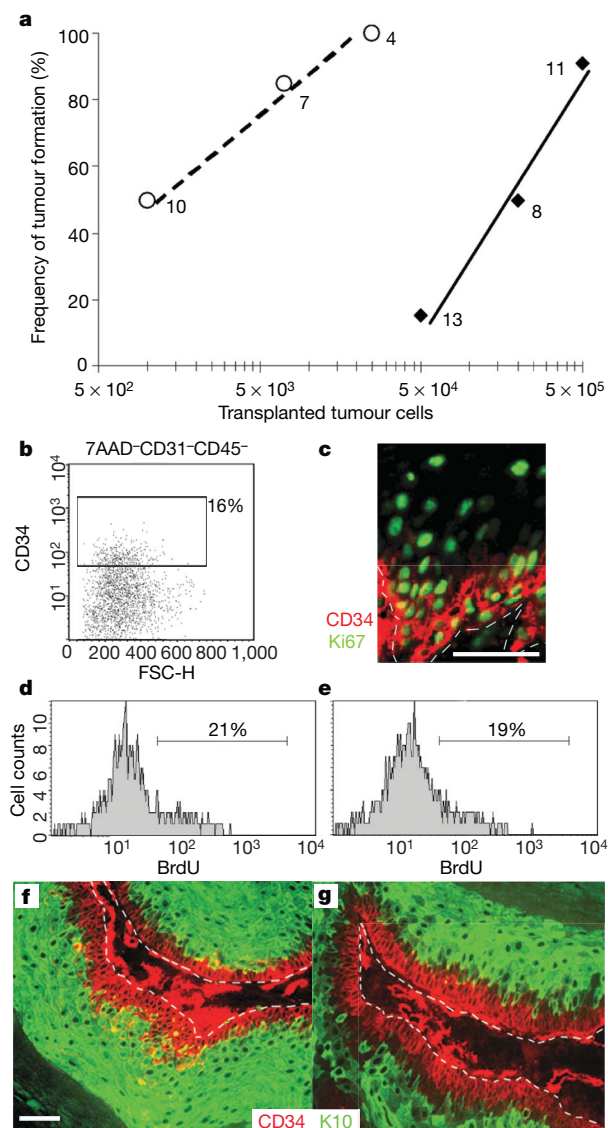

Figure 2 | Cancer stem cells efficiently initiate secondary tumours that recapitulate the organization of the primary tumour. a, Diagram summarizing the frequency of tumour formation in orthotopic tumour transplants using unsorted cells (filled diamonds) or CD34[+] cells (open circles) in varying amounts. The *n* value for each point is shown. **b,** Abundance of CSCs (CD34[+]7AAD[−]CD31[−]CD45[−]) in secondary tumours derived from orthotopic transplantations of CD34[+] cells. **c–e,** Assessment of the distribution of proliferating tumour cells by immunofluorescence analysis for CD34 (red) and Ki67 (green) (**c**) and by short-term BrdU incorporation in CD34[−] cells (**d**) or CD34[+] cells (**e**). **f, g,** Hierarchical organization of primary tumours (**f**) is maintained in secondary tumours (**g**) as determined by immunofluorescence analysis of CD34 (red) and keratin 10 (K10, green). Scale bars, 50 μm.
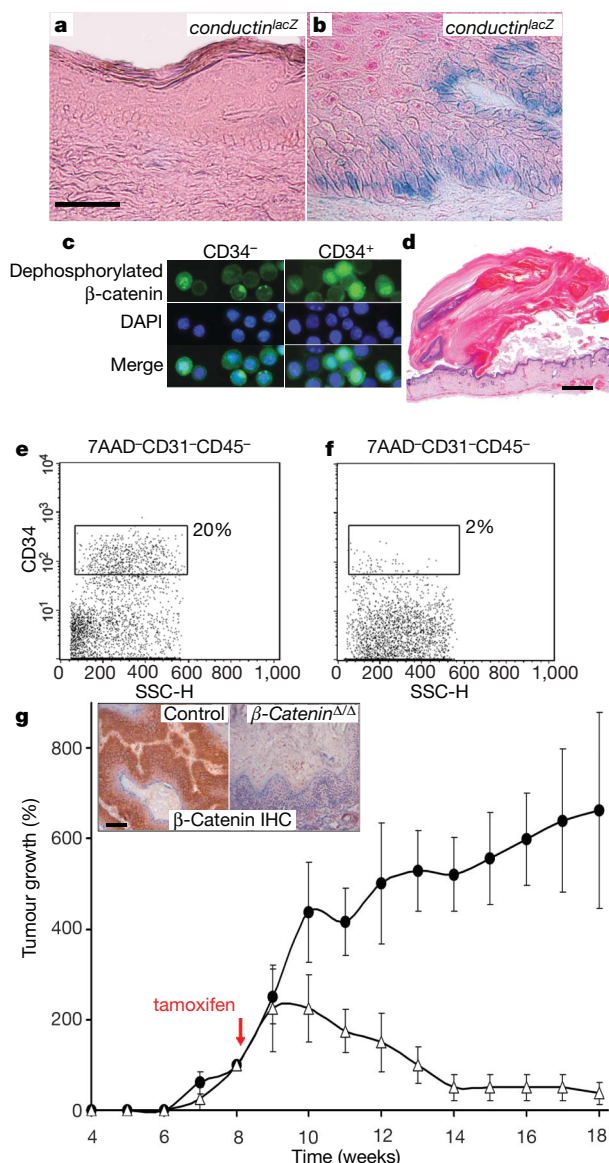
Figure 3 | β-Catenin signalling is essential to maintain skin tumorigenesis.
a, **b**, β-Catenin signalling activity as detected by *lacZ* expression from the
conductin^{lacZ} reporter in normal skin (**a**) and DMBA/TPA-induced tumours
(**b**). **c**, Nuclear, dephosphorylated β-catenin is enriched in CD34$^+$ cells
(right) as detected by immunofluorescence analysis of cytospins after
magnetic sorting. DAPI, 4,6-diamidino-2-phenylindole. **d**, β-Catenin-
negative tumours undergo terminal differentiation after six weeks as
characterized by abundant keratinization. **e**, **f**, Abundance of CD34$^+$ cells in
control tumours (**e**) and β-catenin-deficient tumours (**f**) two weeks after
induction of deletion. **g**, Tumour regression after tamoxifen-induced
β-catenin deletion in established skin tumours (*n* = 6; control (filled circles),
K14-creER^{T2}:β-catenin^{+/Δ}; mutant (open triangles), *K14-creER^{T2}:β-
catenin^{Δ/Δ}*). Tumour numbers at 8 weeks were set to 100%. Error bars show
s.d. Insets: immunohistochemistry (IHC) to detect efficiency of *β-catenin*
gene ablation. Scale bars, 50 μm (**a**, **b**), 400 μm (**d**) and 100 μm (**g**).

wild-type, $17.2 \pm 4.8\%$ (mean ± s.d.); *β-catenin^{Δ/Δ}*, $3.0 \pm 2.0\%$;
Supplementary Fig. 6). This loss of CD34$^+$ cells precedes any
alterations in tumour cell proliferation, apoptosis or senescence,
which remain unaltered at this time point, whereas proliferation is
affected only later when cells differentiate (Supplementary Figs 7 and
8, and data not shown). β-Catenin-deficient tumour cells devoid of
the CD34$^+$ population lose their ability to initiate secondary
tumours, even if transplanted in large numbers (up to 10^6; *n* = 11).
In the reverse experiment we find that the tamoxifen-induced
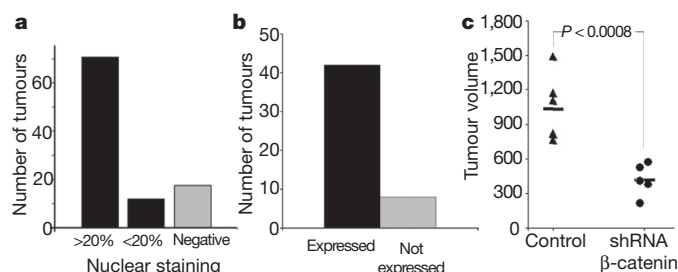expression of an active, non-degradable mutant of β-catenin is

**Figure 4 | Functional importance of β-catenin signalling in human skin
SCCs. a**, **b**, Quantification of nuclear β-catenin (**a**, *n* = 95) and conductin
(**b**, *n* = 50) expression in human skin SCCs. **a**, Tumours displaying nuclear
β-catenin staining were divided into two groups depending on the frequency
of positive cells. **c**, shRNA-mediated knockdown of β-catenin in human
SCC13 cells strongly decreases tumour growth (tumour volume in mm^3) in
xenografts (*P* < 0.0008; Student's *t*-test, *n* = 5; horizontal bars indicate the
means).

sufficient to expand (2.5-fold) the pool of bulge stem cells within
five days (Supplementary Fig. 9; *n* = 8; wild-type, $1.8 \pm 0.2\%$
(mean ± s.d.); *β-catenin^{floxEx3/+}:K14-creER^{T2}*, $4.6 \pm 1.1\%$). We con-
clude that β-catenin has an essential function in the maintenance of
skin CSCs.

Similarly to murine tumours, human skin SCCs show constitutive
activation of β-catenin signalling as revealed by nuclear localization
of β-catenin (Supplementary Fig. 10; *n* = 95). This was observed with
an optimized protocol that detects more abundant signalling (84% of
tumours, Fig. 4a) than previously reported[23–25]. Likewise, expression
of the general Wnt/β-catenin target gene *conductin* is often detected
in human SCCs (Fig. 4b; *n* = 50). Nuclear β-catenin is more frequent
than would be predicted from a small population of CSCs. This
emphasizes that although β-catenin signalling is required for tumor-
igenesis (see also below) it is not strictly a marker of CSCs. Human
skin stem-cell markers have not been described and CD34 is not
expressed in human skin or SCCs, which precluded a CSC analysis
of human tumours. We therefore tested the importance of β-catenin
signalling in human SCC cell lines. We find that short hairpin RNA
(shRNA)-mediated knockdown of β-catenin strongly decreases
tumour growth of human SCC13 cells in xenografts (Fig. 4c and
Supplementary Fig. 11). Residual tumour areas are always positive
for β-catenin, indicating selection for cells that had escaped knock-
down (Supplementary Fig. 11). The main function of β-catenin in
SCCs seems to be lymphoid-enhancer binding factor/T-cell-specific
transcription factor (LEF/TCF)-regulated transcription, because
expression of a dominant-negative mutant of LEF-1 (*ΔNLEF-1*)
decreased tumour formation in human HN-30 SCC xenografts and
in murine orthotopic tumour transplants (Supplementary Fig. 12
and data not shown). This is supported by increased expression of
γ-catenin in murine tumours, which maintains epithelial adhesion in
the absence of β-catenin (Supplementary Fig. 13).

Using transplantation experiments in pre-cancerous murine and
advanced human SCCs we show that β-catenin signalling is essential
for tumorigenesis. In the murine system, this signalling specifically
sustains CSCs, which are necessary for long-term tumour growth.
Potential direct targets of β-catenin signalling that might be respons-
ible for this phenotype include c-*myc*, which was recently shown to be
essential for skin tumorigenesis, but not adult skin homeostasis[26]. As
CSCs are expected to be responsible for tumour recurrence[4], strat-
egies targeting CSCs may lead to more effective therapies. We now
provide evidence that differential requirements for β-catenin signal-
ling in cancer versus normal stem cells exist in skin, which could be
exploited for future therapy.

METHODS SUMMARY
Primary tumour cells, keratinocytes and dermal cells were prepared and trans-
planted, and immunodetection, FACS and *in situ* hybridization were performed.

MACS (Miltenyi) was used to isolate cell populations enriched or depleted for CD34 expression.

The skin-specific ablation of *β-catenin* and the *K14-cre*, *β-catenin*^{floxEx3}, *K14-creER*^{T2}, *conductin*^{lacZ}, *ROSA26-lox-stop-lox-YFP* and *ROSA26-lox-stop-lox-lacZ* alleles have been described[16,21,27–29]. The *K15-creER*^{T2} transgene used 5 kilobases of the K15 promoter, and creER^{T2} was induced by injection of 1 mg of tamoxifen daily for five days. For tumour initiation, mice were treated once with 100 μg of DMBA followed by 12.5 μg of TPA twice per week. Mice expressing the v-*H-ras* transgene (TG.AC)[22], were treated with 5 μg of TPA twice per week only.

Xenografts of human SCC cells were performed in *Nude* mice by using growth-factor-reduced Matrigel with 2×10^6 HN30 or SCC13 cells stably transduced by lentiviruses expressing dominant-negative *ΔNLEF-1* (NM_010703; nucleotides 1086–2183) or shRNAs directed against β-catenin[30] or GFP, respectively.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Ito, M. *et al.* Stem cells in the hair follicle bulge contribute to wound repair but not to homeostasis of the epidermis. *Nature Med.* **11**, 1351–1354 (2005).
2. Al-Hajj, M., Becker, M. W., Wicha, M., Weissman, I. & Clarke, M. F. Therapeutic implications of cancer stem cells. *Curr. Opin. Genet. Dev.* **14**, 43–47 (2004).
3. Reya, T. & Clevers, H. Wnt signalling in stem cells and cancer. *Nature* **434**, 843–850 (2005).
4. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
5. Mackenzie, I. C. Retroviral transduction of murine epidermal stem cells demonstrates clonal units of epidermal structure. *J. Invest. Dermatol.* **109**, 377–383 (1997).
6. Trempus, C. S. *et al.* CD34 expression by hair follicle stem cells is required for skin tumor development in mice. *Cancer Res.* **67**, 4173–4181 (2007).
7. Blanpain, C., Lowry, W. E., Geoghegan, A., Polak, L. & Fuchs, E. Self-renewal, multipotency, and the existence of two cell populations within an epithelial stem cell niche. *Cell* **118**, 635–648 (2004).
8. Tumbar, T. *et al.* Defining the epithelial stem cell niche in skin. *Science* **303**, 359–363 (2004).
9. Liu, Y., Lyle, S., Yang, Z. & Cotsarelis, G. Keratin 15 promoter targets putative epithelial stem cells in the hair follicle bulge. *J. Invest. Dermatol.* **121**, 963–968 (2003).
10. Quintanilla, M., Brown, K., Ramsden, M. & Balmain, A. Carcinogen-specific mutation and amplification of Ha-ras during mouse skin carcinogenesis. *Nature* **322**, 78–80 (1986).
11. Hermann, P. C. *et al.* Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. *Cell Stem Cell* **1**, 313–323 (2007).
12. Gaspar, C. *et al.* Intracellular β-catenin accumulation underlies cancer stemness and metastatic behaviour in an Apc mouse model for mammary tumorigenesis. *Nature.* (submitted).
13. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
14. O'Brien, C. A., Pollett, A., Gallinger, S. & Dick, J. E. A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature* **445**, 106–110 (2007).
15. Morris, R. J. Keratinocyte stem cells: targets for cutaneous carcinogens. *J. Clin. Invest.* **106**, 3–8 (2000).
16. Huelsken, J., Vogel, R., Erdmann, B., Cotsarelis, G. & Birchmeier, W. β-Catenin controls hair follicle morphogenesis and stem cell differentiation in the skin. *Cell* **105**, 533–545 (2001).
17. Merrill, B. J., Gat, U., DasGupta, R. & Fuchs, E. Tcf3 and Lef1 regulate lineage differentiation of multipotent stem cells in skin. *Genes Dev.* **15**, 1688–1705 (2001).
18. Andl, T., Reddy, S. T., Gaddapara, T. & Millar, S. E. Wnt signals are required for the initiation of hair follicle development. *Dev. Cell* **2**, 643–653 (2002).
19. Niemann, C., Owens, D. M., Huelsken, J., Birchmeier, W. & Watt, F. M. Expression of ΔNLef1 in mouse epidermis results in differentiation of hair follicles into squamous epidermal cysts and formation of skin tumours. *Development* **129**, 95–109 (2002).
20. Chan, E. F., Gat, U., McNiff, J. M. & Fuchs, E. A common human skin tumour is caused by activating mutations in β-catenin. *Nature Genet.* **21**, 410–413 (1999).
21. Lustig, B. *et al.* Negative feedback loop of Wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors. *Mol. Cell. Biol.* **22**, 1184–1193 (2002).
22. Leder, A., Kuo, A., Cardiff, R. D., Sinn, E. & Leder, P. v-Ha-ras transgene abrogates the initiation step in mouse skin tumorigenesis: effects of phorbol esters and retinoic acid. *Proc. Natl Acad. Sci. USA* **87**, 9178–9182 (1990).
23. Brasanac, D., Boricic, I., Todorovic, V., Tomanovic, N. & Radojevic, S. Cyclin A and β-catenin expression in actinic keratosis, Bowen's disease and invasive squamous cell carcinoma of the skin. *Br. J. Dermatol.* **153**, 1166–1175 (2005).
24. Doglioni, C. *et al.* Alterations of β-catenin pathway in non-melanoma skin tumors: loss of α-ABC nuclear reactivity correlates with the presence of β-catenin gene mutation. *Am. J. Pathol.* **163**, 2277–2287 (2003).
25. Papadavid, E., Pignatelli, M., Zakynthinos, S., Krausz, T. & Chu, A. C. Abnormal immunoreactivity of the E-cadherin/catenin (α-, β-, and γ-) complex in premalignant and malignant non-melanocytic skin tumours. *J. Pathol.* **196**, 154–162 (2002).
26. Oskarsson, T. *et al.* Skin epidermis lacking the c-Myc gene is resistant to Ras-driven tumorigenesis but can reacquire sensitivity upon additional loss of the *p21*^{Cip1} gene. *Genes Dev.* **20**, 2024–2029 (2006).
27. Harada, N. *et al.* Intestinal polyposis in mice with a dominant stable mutation of the beta-catenin gene. *EMBO J.* **18**, 5931–5942 (1999).
28. Li, M. *et al.* Skin abnormalities generated by temporally controlled RXRα mutations in mouse epidermis. *Nature* **407**, 633–636 (2000).
29. Srinivas, S. *et al.* Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. *BMC Dev. Biol.* **1**, 4 (2001).
30. van de Wetering, M. *et al.* Specific inhibition of gene expression using a stably integrated, inducible small-interfering-RNA vector. *EMBO Rep.* **4**, 609–615 (2003).

Author Contributions H.P., D.K. and T.H. performed research and analysed data; D.H. analysed data; A.C. designed research; D.M., P.C., M.H., D.H. and W.B. contributed vital reagents; I.M. and J.H. designed and performed research, analysed data, and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.H. (joerg.huelsken@epfl.ch).

nature

METHODS

Cell preparation, culture and transplantation. Isolation of cells and transplantations were performed as described[31]. In brief, skins of newborn mice were collected and incubated in 0.25% trypsin in PBS for 16 h. Epidermis was separated from the dermal compartment, chopped and incubated for 1 h at 21 °C in low-calcium medium (LCM). Undissolved tissue clumps were removed by filtration and keratinocytes were plated on collagen (Vitrogen100 collagen)-treated dishes. The dermal compartment was chopped and incubated for 20 min in 0.35% collagenase solution (Sigma) and diluted fivefold with Hanks balanced salt solution (HBSS), and undigested tissue was removed by filtration. The dermal cell suspension was centrifuged at 72g to obtain two fractions: pellet (P1) and supernatant (S1): P1 was resuspended and centrifuged at 10g, and S1 was centrifuged at 220g. The supernatant of P1 and the pellet of S1 were pooled, centrifuged at 72g and washed three times in HBSS. Fibroblasts were resuspended in LCM and plated. Primary adult mouse keratinocytes were prepared as described above except for an initial removal of hair with a depilatory cream (Crème Dépilatoire; Veet). Cells were grown at 5% CO_2 and 34 °C in LCM (2 g l^{-1} KCl, 1 g l^{-1} MgSO$_4$·7H$_2$O, 34 g l^{-1} NaCl, 11 g l^{-1} NaHCO$_3$, 0.7 g l^{-1} NaH$_2$PO$_4$-H$_2$O, 1 g l^{-1} glucose, phenol red (Sigma), non-essential and essential amino acid mix (Invitrogen), vitamins (Invitrogen), antibiotic/antimycotic (Invitrogen), epidermal growth factor (100 ng ml^{-1}), glutamine (Invitrogen), 45 mM CaCl$_2$, 4% fetal calf serum (FCS; Chelex100 treated)).

Tumours were harvested from mice, chopped with a razor blade and incubated in collagenase solution (0.35% in HBSS) for 30–45 min. The suspension was filtered, then washed three times with PBS; cells were resuspended in LCM. Lentiviral infections were performed for 24 h in the presence of Polybrene. Cells were washed three times with PBS and allowed to recover for one or two days after infection.

Orthotropic transplantations were performed on the back of *Nude* mice by using silicon transplantation chambers[31] that were inserted into 2 cm^2 full-thickness excision wounds. Mixtures of freshly prepared tumour cells (10^3 to 10^6), primary keratinocytes (10^6) and dermal cells (8 × 10^5) were injected into these chambers, which were removed one week later, after reconstituted skin had been established.

Studies of human tumour samples were approved by the ethical commission of Lausanne University, and patients' informed consent was obtained. Mouse experiments were performed in accordance with Swiss guidelines and approved by the Veterinarian Office of Vaud, Switzerland.

Immunodetection, FACS and *in situ* hybridization. FACS staining was performed with (3–5) × 10^6 cells in 400 μl of PBS supplemented with 3% FCS and 24.G2 Fc block with the use of APC-coupled CD34 (20 μg ml^{-1}) and biotinylated CD31 and CD45 (Alexa488-streptavidin) antibodies (Pharmingen and eBioscience). Antibodies directed against the following antigens were used for immunodetection on frozen or paraffin tissue sections: CD34 (Pharmingen), Ki67 (Novocastra), BrdU (Sigma), keratin 10 (Covance), β-catenin (BD Transduction Laboratories, Millipore and ref. 16), active caspase-3 (Cell Signaling) and tenascin C (G. Orend and R. Chiquet-Ehrismann). Nuclear β-catenin was detected by using a method described by the Clevers group[32]. *In situ* hybridizations were performed as described previously[16] with the following probes: CD34 (NM_133654, nucleotides (nt) 53–1253), Conductin (NM_015732, nt 395–1039), Conductin (NM_004655, nt 1092–1872), Gas6 (NM_019521, nt 310–1509), ephrin A4 (NM_007910, nt 295–762), Sox9 (NM_011448, nt 10–2530), Runx1 (NM_001111022.1, nt 56–1819) and TCF3 (NM_001079822, nt 1–1030).

LTR and BrdU assays. For analysis of long-term label-retaining (LTR) cells, 10-day-old mice were injected with BrdU (50 mg per kg body weight) every 12 h, with a total of five injections. After 30 days, mice were killed and LTR cells were detected by immunohistochemistry. To achieve short-term labelling of cells undergoing DNA synthesis, adult mice received one intraperitoneal injection of BrdU at 100 mg per kg body weight 2 h before being killed.

31. Lichti, U. *et al. In vivo* regulation of murine hair growth: insights from grafting defined cell populations onto nude mice. *J. Invest. Dermatol.* **101**, 124–129 (1993).
32. Batlle, E. *et al.* β-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* **111**, 251–263 (2002).

LETTERS

Dendritic cell PAR1–S1P3 signalling couples coagulation and inflammation

Frank Niessen[1], Florence Schaffner[1], Christian Furlan-Freguia[1], Rafal Pawlinski[1], Gourab Bhattacharjee[1], Jerold Chun[2], Claudia K. Derian[4], Patricia Andrade-Gordon[4], Hugh Rosen[1,3] & Wolfram Ruf[1]

Defining critical points of modulation across heterogeneous clinical syndromes may provide insight into new therapeutic approaches. Coagulation initiated by the cytokine-receptor family member known as tissue factor is a hallmark of systemic inflammatory response syndromes in bacterial sepsis and viral haemorrhagic fevers[1,2], and anticoagulants can be effective in severe sepsis with disseminated intravascular coagulation[3]. The precise mechanism coupling coagulation and inflammation remains unresolved[4–7]. Here we show that protease-activated receptor 1 (PAR1) signalling sustains a lethal inflammatory response that can be interrupted by inhibition of either thrombin or PAR1 signalling. The sphingosine 1-phosphate (S1P) axis is a downstream component of PAR1 signalling, and by combining chemical and genetic probes for S1P receptor 3 (S1P3) we show a critical role for dendritic cell PAR1–S1P3 cross-talk in regulating amplification of inflammation in sepsis syndrome. Conversely, dendritic cells sustain escalated systemic coagulation and are the primary hub at which coagulation and inflammation intersect within the lymphatic compartment. Loss of dendritic cell PAR1–S1P3 signalling sequesters dendritic cells and inflammation into draining lymph nodes, and attenuates dissemination of interleukin-1β to the lungs. Thus, activation of dendritic cells by coagulation in the lymphatics emerges as a previously unknown mechanism that promotes systemic inflammation and lethality in decompensated innate immune responses.

Disseminated intravascular coagulation and systemic inflammation are signs of excessive activation of the innate immune system. Both are attenuated by genetic reduction of tissue factor and its protease ligand coagulation factor VIIa, leading to improved survival in endotoxaemia[6,8]. In a model of severe, but not completely lethal lipopolysaccharide (LPS) challenge[9], we show that PAR1 deficiency protects mice from lethality (Fig. 1a). PAR1[−/−] mice initially developed elevated inflammation and coagulation markers indistinguishable from the wild type (Fig. 1b, c). Unlike the wild type, PAR1[−/−] mice progressively resolved systemic inflammation beginning at 12 h. To address whether coagulation amplifies inflammation
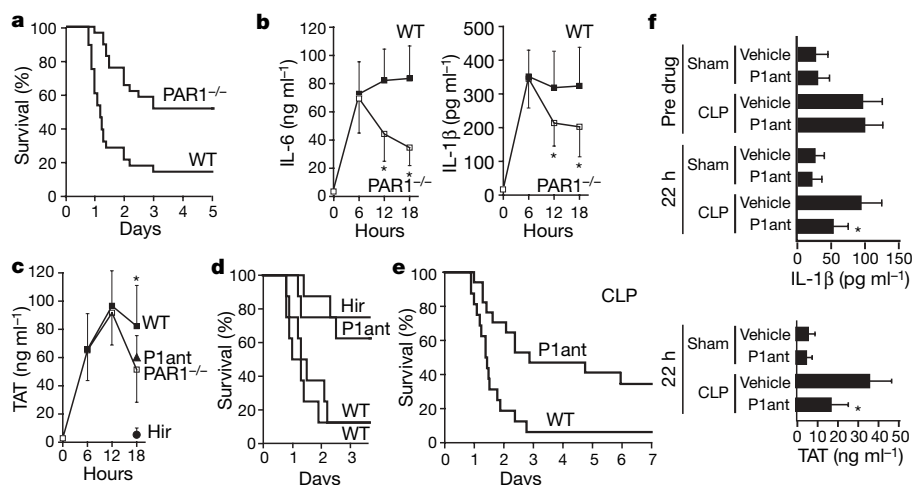

Figure 1 | Coagulation amplifies inflammation and lethality through PAR1 signalling. a, Survival advantage of PAR1[−/−] mice in 90% lethal LPS challenge induced by intraperitoneal injection of 8 mg kg[−1] LPS (summary of three independent experiments, $n \geq 28$ per genotype, PAR1[−/−] survival advantage for each individual experiment, $P < 0.05$). **b,** Reduced late-stage inflammation in PAR1[−/−] mice documented by IL-6 and IL-1β levels (mean ± s.d., $n = 18$ per group, asterisks indicate groups that are different from the wild type (WT), $P < 0.05$). **c,** TAT levels in wild-type and PAR1[−/−] mice, or wild-type mice treated at 10 h with PAR1 antagonist RWJ58259 (P1ant) or the thrombin inhibitor hirudin (Hir). **d,** Intervention with PAR1 antagonist or hirudin improves survival, similarly to PAR1 deficiency ($n = 8$ per group, $P < 0.02$ relative to wild-type control). **e,** Intervention with PAR1 antagonist (5 mg kg[−1] at 16 and 22 h) improves survival of wild-type mice in the CLP model ($n = 16$ per group, $P < 0.005$, pooled data from two experiments, survival advantage in each experiment for drug-treated group, $P < 0.05$). **f,** Intervention with PAR1 antagonist improves inflammation and coagulation markers. Sham- (surgery without CLP) or CLP-operated mice received PAR1 antagonist or vehicle at 16 h. Pre-drug 16-h retro-orbital and terminal 22-h inferior vena cava blood samples were collected (mean ± s.d., $n = 3$ for sham, $n = 9$ for CLP drug- or vehicle-treated animals, asterisks indicate groups that are different compared with the control, $P < 0.02$ (TAT), $P = 0.001$ (IL-1β)).

[1]Departments of Immunology, [2]Molecular Biology, and [3]Chemical Physiology, The Scripps Research Institute, La Jolla, California 92037, USA. [4]Johnson & Johnson PRD, Spring House, Pennsylvania 19477, USA.

through PAR1 signalling at this stage of the model, wild-type mice were treated 10 h after challenge with either the thrombin inhibitor hirudin or a PAR1-selective antagonist. Both pharmacological interventions protected mice from lethality (Fig. 1d). Like PAR1 deficiency, PAR1 antagonism also attenuated late-stage coagulation (Fig. 1c), indicating that continuing activation of coagulation and inflammation are coupled by a common upstream mechanism.

The implications of these findings for sepsis therapy were further evaluated in the caecal ligation and puncture (CLP) model. Wild-type mice were treated with PAR1 antagonist after development of severe sepsis symptoms 16 h after surgery. Blocking PAR1 signalling attenuated lethality and interrupted ongoing inflammation as well as coagulation (Fig. 1e, f). These data show that intervention in the PAR1 pathway improves sepsis outcome, and indicate that appropriate LPS challenge adequately models mechanisms by which PAR1 contributes to sepsis lethality.

Multiplex cytokine profiles confirmed that early, LPS-induced inflammatory responses of the innate immune system were unaltered in PAR1[−/−] mice (Fig. 2a). In contrast, PAR1 deficiency attenuated broadly inflammatory parameters 18 h after LPS challenge (Fig. 2b). Delayed intervention with coagulation inhibitor or PAR1 antagonist in wild-type mice recapitulated the reduction in cytokine levels of PAR1[−/−] mice. Thus, coagulation amplifies inflammation, and pharmacological blockade of PAR1 is sufficient to rebalance exacerbated systemic inflammation.

S1P signalling has diverse roles in inflammation and immunity[10], and PAR1 and S1P signalling are coupled in endothelial cells[11,12]. Inflammatory exacerbation was attenuated in mice lacking sphingosine kinase 1 (SphK1), comparable to that seen in PAR1[−/−] mice (Fig. 2b). To show that PAR1 signalling operated upstream of S1P signalling, S1P receptors of PAR1[−/−] mice were directly stimulated with an agonist (AAL) for both S1P1 and S1P3, or a selective agonist (AUY) for S1P1. In control experiments, both agonists induced S1P1-dependent sequestration of T cells by > 90% for 12 h, confirming equivalent potency at the dose that was given as a single bolus injection 10 h after LPS challenge. S1P1/3 agonism by AAL, but not selective S1P1 activation by AUY, completely reversed the protection of PAR1[−/−] mice from exacerbated inflammation (Fig. 2c). S1P3[−/−] mice showed attenuated cytokine levels similar to PAR1[−/−] mice, but the S1P1/3 agonist AAL had no pro-inflammatory activity in S1P3[−/−] mice, confirming that inflammation is specifically amplified by S1P3 signalling.

Although we initially assumed that vascular cells initiated this pro-inflammatory signalling cross-talk, reconstitution of PAR1[−/−] mice with wild-type bone marrow was sufficient to restore inflammation in PAR1[−/−] mice (Fig. 3a). The cytokine profile indicated T- and dendritic cell activation, but T-cell counts in late-stage LPS challenge were similar between wild-type and S1P3[−/−], PAR1[−/−] or PAR1-antagonist-treated mice (Supplementary Fig. 2a, b). To address the role of dendritic cells, wild-type bone-marrow-derived dendritic cells (Supplementary Fig. 2c, d) were adoptively transferred 24 h before challenge. This restored late-stage inflammation in PAR1[−/−], SphK1[−/−] or S1P3[−/−] mice, but did not further amplify inflammation in wild-type mice, showing that adoptive transfer of dendritic cells per se did not exacerbate inflammation (Fig. 3a). In addition, PAR1[−/−] or S1P3[−/−] dendritic cells did not increase systemic inflammation after adoptive transfer into S1P3[−/−] or SphK1[−/−] mice. Thus, each component of the pathway must be present on the dendritic cell to restore systemic inflammation, indicating that PAR1, SphK1 and S1P3 are coupled in an autocrine, rather than paracrine, pathway.

Dendritic cells have important roles in the activation and resolution of innate immune responses[13,14], and inflammation mobilizes migratory dendritic cells from the periphery to lymph nodes[15]. Because adoptively transferred dendritic cells were recruited to lymph nodes (Supplementary Fig. 3a), we analysed local inflammation in draining mesenteric lymph nodes that are downstream of the severe insult of intraperitoneal LPS administration. Surprisingly, PAR1[−/−], SphK1[−/−] and S1P3[−/−] mice had increased IL-1β levels in mesenteric lymph nodes, but IL-1β levels were conversely reduced in the lungs, which represent the first microvascular bed downstream of lymphatic drainage into the thoracic duct (Fig. 3b). In support of the concept that loss of PAR1–S1P3 signalling specifically contained inflammation in draining lymph nodes, IL-1β levels in the spleens were unchanged in all knockout strains. Furthermore, wild-type mice receiving S1P3[−/−] dendritic cells had increased local inflammation in mesenteric lymph nodes, whereas transfer of wild-type dendritic cells into S1P3[−/−] mice promoted dissemination of inflammation to the lungs. Importantly, PAR1 antagonism in the CLP model induced sequestration of inflammation to draining mesenteric lymph nodes and prevented dissemination of inflammation to the lungs (Fig. 3b). This shows that the PAR1 pathway can be targeted successfully in polymicrobial sepsis syndrome.

Elevated lymph node IL-1β levels were strictly correlated with increased mesenteric lymph node sizes in the LPS and CLP model
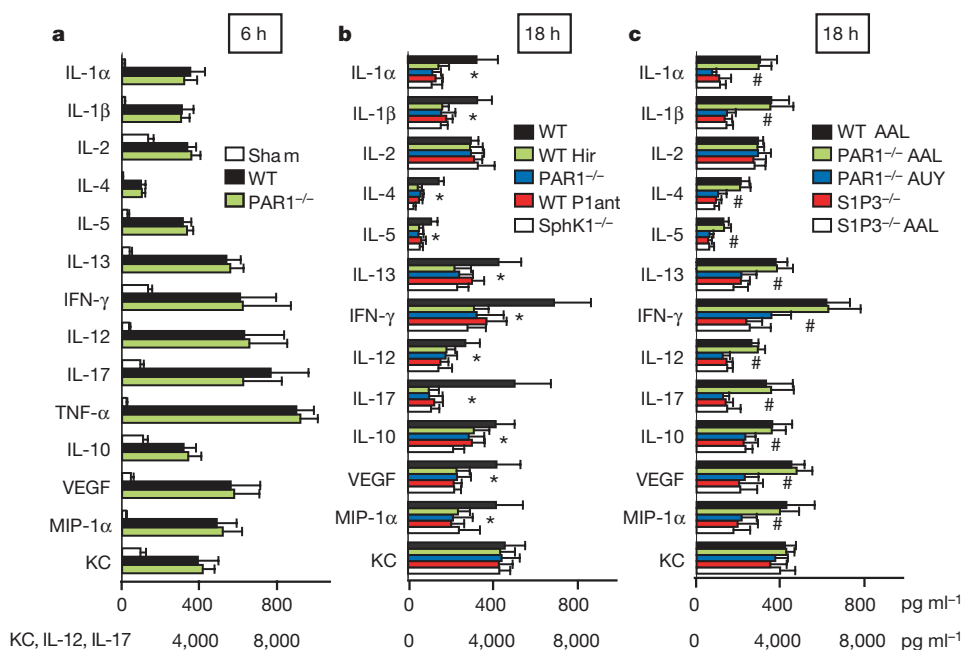

Figure 2 | PAR1 amplifies inflammation through SphK1–S1P3 signalling crosstalk. a, Cytokine multiplex analysis[9] documents that PAR1[−/−] mice have no alterations in the inflammatory response 6 h after LPS challenge. **b**, Thrombin inhibition, PAR1 antagonism and SphK1 deficiency mimic the broad attenuation of inflammation seen in PAR1[−/−] mice 18 h after challenge. Note that some cytokines (KC (CXCL1) and IL-2) remain elevated independent of PAR1 signalling. **c**, Intervention at 10 h with the S1P1/3 agonist AAL, but not with the selective S1P1 agonist AUY, reverses attenuated inflammation in PAR1[−/−], but not S1P3[−/−] mice. (mean ± s.d., n = 7–16, asterisks indicate groups that are different compared with the wild type, hash symbols indicate groups that are different compared with wild type AAL or PAR1[−/−] AAL, P < 0.05 by ANOVA).

(Supplementary Fig. 3b, d), confirming at the macroscopic level that local sequestration of inflammation is controlled by dendritic cells. Because S1P3 signalling enhances dendritic cell motility[16,17], we addressed whether loss of PAR1–S1P3 signalling led to accumulation of dendritic cells in lymph nodes. To quantify accurately dendritic cells, without influence from lymph node size variations, wild-type or PAR1[−/−] dendritic cells tagged with green fluorescent protein (GFP) were adoptively transferred into S1P3[−/−] mice. Cell density of mesenteric lymph nodes 18 h after LPS challenge was not changed by adoptive transfer, but PAR1[−/−] dendritic cells were enriched two- to threefold compared with the wild type (Fig. 3c). Thus, loss of PAR1 signalling sequesters dendritic cells into draining lymph nodes.

IL-1β release from dendritic cells is triggered by ATP-dependent signalling of the P2X7 receptor, which leads to activation of caspase-1 in an adaptor complex, the 'inflammasome'[18]. Adoptive transfer of P2X7[−/−] dendritic cells into S1P3[−/−] mice failed to increase lung IL-1β levels (Fig. 3b). To further corroborate that dendritic cell S1P signalling regulates IL-1β dissemination, S1P3[−/−] or wild-type mice were adoptively transferred with PAR1[−/−] dendritic cells. In S1P3[−/−] mice, PAR1[−/−] dendritic cells are the only target for S1P3 agonism by AAL delivered 10 h after LPS challenge. Adoptive transfer

of PAR1[−/−] dendritic cells into S1P3[−/−] mice did not induce dissemination of IL-1β to the lungs, unless PAR1[−/−] dendritic cells were stimulated with the S1P3 agonist. Conversely, PAR1[−/−] dendritic cells adoptively transferred into wild-type mice increased local inflammation specifically in mesenteric, but not inguinal, lymph nodes, and S1P agonism with AAL reversed increased lymph node IL-1β levels (Fig. 3b and Supplementary Fig. 3c). Thus, PAR1–S1P3 signalling is coupled to activation of the dendritic cell inflammasome to promote systemic dissemination of inflammation in sepsis.

Dendritic cells are known to express tissue factor[19], and therapeutic efficacy of hirudin implicated thrombin in a feedback loop that regulates dendritic cell function. Fibrin was detected in lymphatic sinuses that were identified by LYVE-1 staining, confirming thrombin generation in this unexpected location of mesenteric lymph nodes (Fig. 4a). In addition, a protective thrombin inhibitor dose reduced lymphatic fibrin deposition. Although fibrin deposition was sparse in S1P3[−/−] lymph nodes, PAR1[−/−] lymph nodes showed abundant fibrin throughout the parenchyma that contrasted with the predominant wild-type staining within lymphatic ducts. These data indicated that wild-type mice lost the ability to sequester coagulation activation to draining lymph nodes.
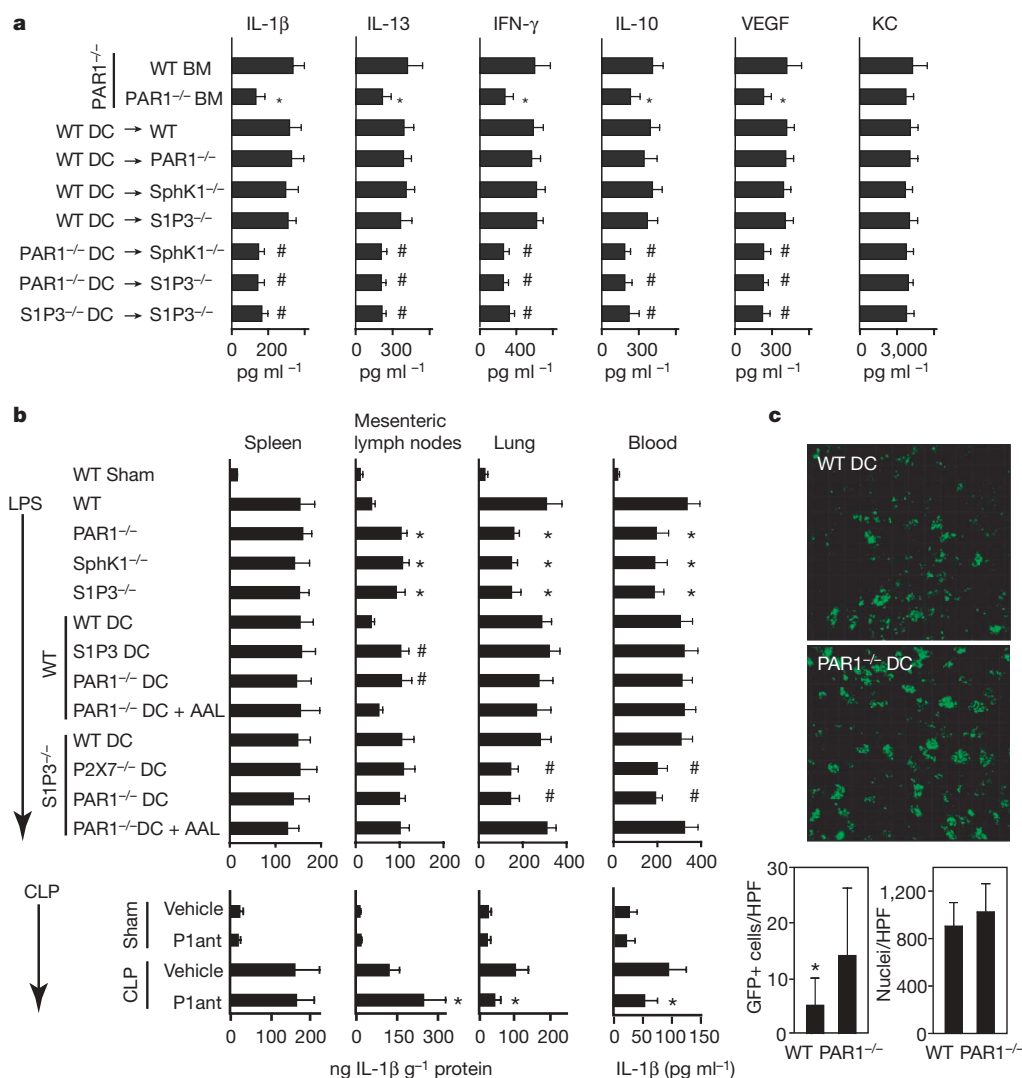

Figure 3 | Dendritic cell PAR1–SphK1–S1P3 signalling controls dissemination of inflammation from the lymphatics. a, Wild-type bone marrow (BM) chimaeras of PAR1[−/−] mice lose protection from systemic inflammation. PAR1[−/−] or wild-type GFP[+] bone marrow resulted in similar reconstitution of PAR1[−/−] mice (peripheral blood GFP[+] cells: CD4[+], 44 ± 15 versus 51 ± 15%; CD8[+], 31 ± 19 versus 37 ± 19%; CD11c[+], 62 ± 21 versus 72 ± 5%; B220[+], 93 ± 2 versus 92 ± 2%). Wild-type dendritic cells (DC) adoptively transferred by tail vein injection 24 h before challenge reverse attenuated inflammation in protected strains (mean ± s.d., n = 4–6 mice per group, asterisks indicate groups that are different compared with the wild-type bone marrow, hash symbols indicate groups that are different from the wild-type dendritic cell group, P < 0.05 by ANOVA). **b**, Blood and tissue IL-1β levels of spleens, mesenteric lymph nodes and lungs measured 18 h after LPS challenge with AAL administration at 10 h, or 22 h after CLP as described in Fig. 1f (mean ± s.d., n = 4–6 mice per group, asterisks indicate groups that are different from challenged wild type without drug, hash symbols indicate groups that are different compared with animals receiving wild-type dendritic cells, P < 0.05 by ANOVA). **c**, Retention of PAR1[−/−] dendritic cells in S1P3[−/−] lymph nodes during late-stage endotoxaemia. Typical experiment of quantitative image analysis (two independent experiments with n = 3, *P < 0.05) and views of three-dimensional reconstructions of 30-μm sections from mesenteric lymph nodes. The grid lines are 20 μm apart; HPF, high power field.

Thrombin-antithrombin (TAT) levels in the periphery further support the concept that dendritic cells initiate dissemination of intravascular coagulation from the lymphatics. TAT levels in peripheral blood samples were close to baseline in S1P3$^{-/-}$ mice (Fig. 4b), although several independent parameters indicated normal LPS-induced tissue-factor expression in S1P3$^{-/-}$ mice (Supplementary Fig. 4). PAR1$^{-/-}$ or SphK1$^{-/-}$ mice had higher peripheral TAT levels compared with S1P3$^{-/-}$ mice. The precise mechanism for the complete loss of late-stage coagulation activation in S1P3$^{-/-}$ mice requires further study. However, adoptive transfer of wild-type dendritic cells restored peripheral TAT levels in S1P3$^{-/-}$, PAR1$^{-/-}$ and SphK1$^{-/-}$ mice to levels of wild-type mice. In addition, PAR1$^{-/-}$ dendritic cells increased coagulation in S1P3$^{-/-}$ mice to levels observed in SphK1$^{-/-}$ or PAR1$^{-/-}$ mice. Thus, dendritic cells are the primary source and determine the extent of disseminated coagulation activation during inflammatory exacerbation.

Attenuation of coagulation is not required for survival in the LPS model[9] and improved survival of S1P3$^{-/-}$ and SphK1$^{-/-}$ mice indicated that dissemination of inflammation caused lethality (Fig. 4c).

To specifically show that PAR1 signalling on dendritic cells drives lethality, wild-type or PAR1$^{-/-}$ dendritic cells were adoptively transferred into SphK1$^{-/-}$ mice. Consistent with changes in inflammatory cytokines, wild-type, but not PAR1$^{-/-}$ dendritic cells reversed the survival benefit of SphK1$^{-/-}$ mice (Fig. 4d). In contrast, adoptive transfer of the same number of wild-type, bone-marrow-derived macrophages[9] into SphK1$^{-/-}$ mice had no adverse effect on survival, demonstrating specificity. Furthermore, administration of the direct S1P1/3 agonist (AAL) reversed the therapeutic benefit of PAR1 antagonism in wild-type mice (Fig. 4e), despite potential vascular protective S1P1 agonistic activity[11].

To show directly that dendritic cell S1P signalling initiates lethality, S1P3$^{-/-}$ mice were adoptively transferred with PAR1$^{-/-}$ dendritic cells. This had no adverse effect on survival of S1P3$^{-/-}$ mice, but stimulation of S1P3 with AAL on adoptively transferred PAR1$^{-/-}$ dendritic cells was sufficient to induce lethality (Fig. 4f). Consistent with detrimental effects of S1P3 signalling on both vascular and dendritic cells, S1P3$^{-/-}$ mice showed delayed lethality in the CLP model (see Supplementary Figs 5, 6 for further discussion of vascular
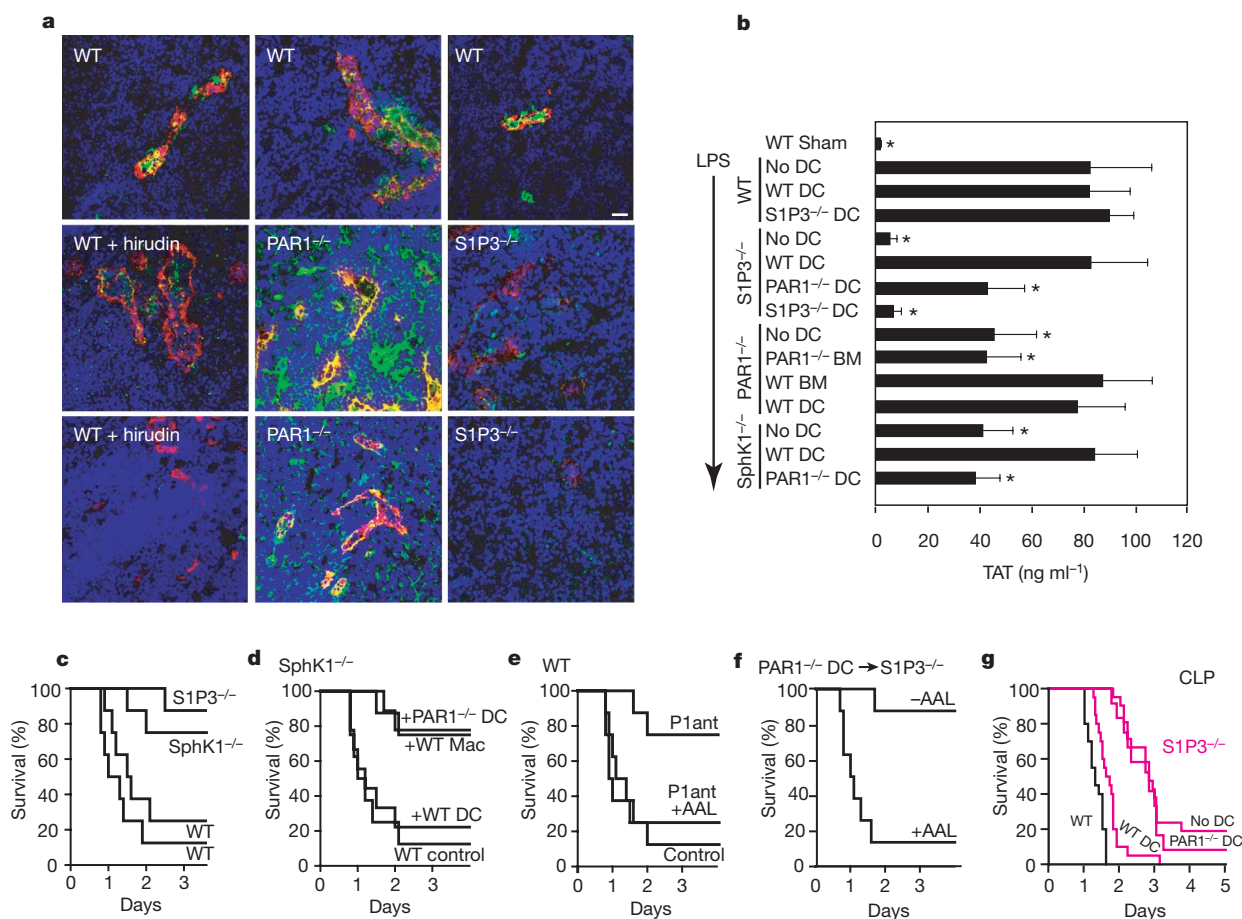

Figure 4 | Dendritic cell PAR1–S1P3 signalling promotes disseminated intravascular coagulation and lethality. a, Fibrin staining (green) of mesenteric lymph nodes from wild-type, S1P3$^{-/-}$ or PAR1$^{-/-}$ mice 18 h after LPS challenge with hirudin administration at 10 h. Sections were counterstained for LYVE-1 (red) to visualize lymphatic ducts. **b,** Dendritic cells are responsible for exacerbated disseminated intravascular coagulation. The graph shows TAT levels at 18 h in peripheral venous blood samples of LPS-challenged mice that were adoptively transferred with dendritic cells from the indicated knockouts or the indicated bone-marrow chimaeras in PAR1$^{-/-}$ mice (mean ± s.d., $n = 4$–10, $P < 0.05$ versus respective wild-type dendritic cell or wild-type bone marrow by ANOVA). **c–f,** Dendritic cell S1P signalling promoted lethality. **c,** Protection of S1P3$^{-/-}$ and SphK1$^{-/-}$ mice from LPS challenge ($n = 8$ per group, $P < 0.01$ relative to wild type). **d,** Adoptive transfer of wild-type, but not PAR1$^{-/-}$ dendritic cells reverses improved survival of SphK1$^{-/-}$ mice ($P = 0.008$, $n = 9$ per group) and lethal

effects are specific for transfer of dendritic cells, but not macrophages (Mac) ($P = 0.003$, $n = 8$ per group). **e,** S1P3 agonism with AAL reverses survival of PAR1 antagonist-treated mice ($P = 0.004$ for PAR1 antagonist versus wild-type control, $P = 0.016$ for PAR1 antagonist ± AAL, $n = 8$ per group). **f,** Dendritic cell S1P agonism with AAL at 10 h promotes LPS-induced lethality in S1P3$^{-/-}$ mice adoptively transferred with PAR1$^{-/-}$ dendritic cells ($P = 0.0008$, $n = 8$ per group). **g,** Adoptively transferred wild-type, but not PAR1$^{-/-}$ dendritic cells promote lethality in S1P3$^{-/-}$ mice in the CLP model ($n \geq 20$ per group, pooled data from two independent experiments; $n = 12$ for S1P3$^{-/-}$ + PAR1$^{-/-}$ dendritic cells. Survival advantage: S1P3$^{-/-}$ versus wild type, $P = 0.0001$; S1P3$^{-/-}$ + wild-type dendritic cells versus wild type, $P = 0.0001$; S1P3$^{-/-}$ + PAR1$^{-/-}$ dendritic cells versus wild type, $P = 0.0001$; S1P3$^{-/-}$ versus S1P3$^{-/-}$ + wild-type dendritic cells, $P = 0.0001$; S1P3$^{-/-}$ + PAR1$^{-/-}$ dendritic cells versus S1P3$^{-/-}$ + wild-type dendritic cells, $P = 0.002$; S1P3$^{-/-}$ versus S1P3$^{-/-}$ + PAR1$^{-/-}$ dendritic cells, $P = 0.5$).

roles of S1P and PAR signalling). Importantly, adoptive transfer of wild-type, but not PAR1$^{-/-}$ dendritic cells impaired survival of S1P3$^{-/-}$ mice in the CLP model (Fig. 4g), confirming the crucial role of dendritic cell PAR1 signalling in inducing sepsis lethality.

These experiments uncover a new link between protease and sphingolipid signalling in the immune system, and position coagulation signalling in the lymphatics upstream of continuing disseminated intravascular coagulation, lung injury and peripheral vascular dysfunction (Supplementary Fig. 1). Dendritic cells emerge as both the target for pro-inflammatory protease signalling and the origin of coagulation activation. This provides a surprisingly simple mechanism for the intimate coupling of exacerbated inflammation and coagulation in deregulated innate immune responses. Although dendritic cells are known to regulate sepsis outcome[20–22], the results presented emphasize the role of dendritic cells as promoters of systemic inflammation and sepsis lethality[23]. Pharmacological blockade of PAR1 is sufficient to interrupt exacerbated systemic dissemination of inflammation and coagulation, and to improve survival in sepsis. The requirement to block thrombin in the lymphatics for anti-inflammatory benefit may explain the limited clinical efficacy of intravenously administered plasma protease inhibitors, such as antithrombin[3]. Importantly, blockade of dendritic cell PAR1–S1P3 signalling in the lymphatics attenuates systemic inflammation by containing inflammation to draining lymph nodes downstream of severe inflammation or infection. We propose that intervention in the coagulation–PAR1–S1P3 axis may avoid immune paralysis and compromised host defence that has hampered therapeutic development of direct anti-inflammatory strategies in decompensated infectious diseases.

METHODS SUMMARY

Under approved protocols, pathogen-free C57BL/6-wild-type, PAR1$^{-/-}$ (ref. 24), S1P3$^{-/-}$ (ref. 25), SphK1$^{-/-}$ (ref. 26), or P2X7 (ref. 27) mice were challenged with LD90, $\sim$8 mg kg^{-1} LPS[9]. PAR1 antagonist RWJ58259 (5 mg kg^{-1}), S1P1/S1P3 agonist AAL(R) (2-amino-4-(4-heptyloxyphenyl)-2-methylbutanol) (0.5 mg kg^{-1})[28], S1P1 agonist AUY (AUY954) (1 mg kg^{-1})[29] or thrombin inhibitor hirudin (120 mg kg^{-1} lepirudin, 1 mg kg^{-1} PEG hirudin) were bolus-injected intravenously. For CLP, mice were anaesthetized (ketamine/xylazine, 100/8 mg kg^{-1}) to ligate 1 cm of caecum for two punctures (21G needles), confirmed to be patent by extrusion of stool. After repositioning of the caecum and wound closure, mice received subcutaneously 0.5 ml of warm saline for recovery. Bone marrow chimaeras were generated by reconstituting lethally irradiated PAR1$^{-/-}$ mice. Bone-marrow-derived dendritic cells were expanded for 7 days[30], positively selected for CD11c$^+$ with paramagnetic beads, and 1×10^6 cells per mouse were injected intravenously 24 h before challenge. Cytokines and TAT were determined in vena cava inferior plasma. Tissue IL-1β levels were measured in cleared extracts (30 mM Tris pH 7.4, 150 mM NaCl, 1% Triton-X100, 2 mM CaCl$_2$ and 2 mM MgCl$_2$, protease inhibitor cocktail) from homogenized lymph nodes, spleens and lungs with normalization for total protein concentration. Immunohistochemistry used frozen 5-μm methanol-fixed sections permeabilized with 0.1% Triton for staining with antifibrin antibody (Nordic), anti-LYVE-1 (Reliatech) and counterstaining with DAPI (Vector). Z-stacks of 30-μm frozen sections were taken with the BioRad Radiance 2100 Rainbow laser scanning confocal microscope (Nikon TE 2000-U). Reconstruction of a three-dimensional z-stack and quantitative analysis was performed with the IMARIS software (Bitplane). Means $\pm$ s.d. are given and normally distributed data were evaluated by analysis of variance (ANOVA) followed by Fisher's PLSD, or by Kruskal–Wallis followed by the Mann–Whitney U-test. Survival advantages were analysed by Kaplan–Meier curves and the log rank test with Bonferroni correction as needed.

1. Esmon, C. T. Interactions between the innate immune and blood coagulation systems. *Trends Immunol.* 25, 536–542 (2004).
2. Ruf, W. Emerging roles of tissue factor in viral hemorrhagic fever. *Trends Immunol.* 25, 461–464 (2004).
3. Opal, S. M. The nexus between systemic inflammation and disordered coagulation in sepsis. *J. Endotoxin Res.* 10, 125–129 (2004).
4. Aird, W. C. The role of the endothelium in severe sepsis and multiple organ dysfunction syndrome. *Blood* 101, 3765–3777 (2003).
5. Taylor, F. B. Jr. Staging of the pathophysiologic responses of the primate microvasculature to *Escherichia coli* and endotoxin: examination of the elements of the compensated response and their links to the corresponding uncompensated lethal variants. *Crit. Care Med.* 29, S78–S89 (2001).
6. Pawlinski, R. *et al.* Role of tissue factor and protease activated receptors in a mouse model of endotoxemia. *Blood* 103, 1342–1347 (2003).
7. Camerer, E. *et al.* Roles of protease-activated receptors in a mouse model of endotoxemia. *Blood* 107, 3912–3921 (2006).
8. Xu, H., Ploplis, V. A. & Castellino, F. J. A coagulation factor VII deficiency protects against acute inflammatory responses in mice. *J. Pathol.* 210, 488–496 (2006).
9. Ahamed, J. *et al.* Regulation of macrophage procoagulant responses by the tissue factor cytoplasmic domain in endotoxemia. *Blood* 109, 5251–5259 (2007).
10. Rosen, H. & Goetzl, E. J. Sphingosine 1-phosphate and its receptors: an autocrine and paracrine network. *Nature Rev. Immunol.* 5, 560–570 (2005).
11. Feistritzer, C. & Riewald, M. Endothelial barrier protection by activated protein C through PAR1-dependent sphingosine 1-phosphate receptor-1 crossactivation. *Blood* 105, 3178–3184 (2005).
12. Singleton, P. A. *et al.* Attenuation of vascular permeability by methylnaltrexone: role of mOP-R and S1P3 transactivation. *Am. J. Respir. Cell Mol. Biol.* 37, 222–231 (2007).
13. Shortman, K. & Naik, S. H. Steady-state and inflammatory dendritic-cell development. *Nature Rev. Immunol.* 7, 19–30 (2007).
14. Steinman, R. M. & Banchereau, J. Taking dendritic cells into medicine. *Nature* 449, 419–426 (2007).
15. Randolph, G. J., Angeli, V. & Swartz, M. A. Dendritic-cell trafficking to lymph nodes through lymphatic vessels. *Nature Rev. Immunol.* 5, 617–628 (2005).
16. Czeloth, N. *et al.* Sphingosine-1-phosphate mediates migration of mature dendritic cells. *J. Immunol.* 175, 2960–2967 (2005).
17. Maeda, Y. *et al.* Migration of CD4 T cells and dendritic cells toward sphingosine 1-phosphate (S1P) is mediated by different receptor subtypes: S1P regulates the functions of murine mature dendritic cells via S1P receptor type 3. *J. Immunol.* 178, 3437–3446 (2007).
18. Ferrari, D. *et al.* The P2X7 receptor: a key player in IL-1 processing and release. *J. Immunol.* 176, 3877–3883 (2006).
19. Baroni, M. *et al.* Stimulation of P2 (P2X7) receptors in human dendritic cells induces the release of tissue factor-bearing microparticles. *FASEB J.* 21, 1926–1933 (2007).
20. Efron, P. A. *et al.* Characterization of the systemic loss of dendritic cells in murine lymph nodes during polymicrobial sepsis. *J. Immunol.* 173, 3035–3043 (2004).
21. Scumpia, P. O. *et al.* CD11c$^+$ dendritic cells are required for survival in murine polymicrobial sepsis. *J. Immunol.* 175, 3282–3286 (2005).
22. Fujita, S. *et al.* Regulation of dendritic cells act as regulators of acute lethal systemic inflammatory response. *Blood* 107, 3656–3664 (2006).
23. Ohteki, T. *et al.* Essential roles of DC-derived IL-15 as a mediator of inflammatory responses *in vivo. J. Exp. Med.* 203, 2329–2338 (2006).
24. Damiano, B. P. *et al.* Cardiovascular responses mediated by protease-activated receptor- 2 (PAR-2) and thrombin receptor (PAR-1) are distinguished in mice deficient in PAR-2 or PAR-1. *J. Pharmacol. Exp. Ther.* 288, 671–678 (1999).
25. Ishii, I. *et al.* Selective loss of sphingosine 1-phosphate signaling with no obvious phenotypic abnormality in mice lacking its G protein-coupled receptor, LP(B3)/EDG-3. *J. Biol. Chem.* 276, 33697–33704 (2001).
26. Allende, M. L. *et al.* Mice deficient in sphingosine kinase 1 are rendered lymphopenic by FTY720. *J. Biol. Chem.* 279, 52487–52492 (2004).
27. Labasi, J. M. *et al.* Absence of the P2X7 receptor alters leukocyte function and attenuates an inflammatory response. *J. Immunol.* 168, 6436–6445 (2002).
28. Don, A. S. *et al.* Essential requirement for sphingosine kinase 2 in a sphingolipid apoptosis pathway activated by FTY720 analogues. *J. Biol. Chem.* 282, 15833–15842 (2007).
29. Pan, S. *et al.* A monoselective sphingosine-1-phosphate receptor-1 agonist prevents allograft rejection in a stringent rat heart transplantation model. *Chem. Biol.* 13, 1227–1234 (2006).
30. Lutz, M. B. *et al.* An advanced culture method for generating large quantities of highly pure dendritic cells from mouse bone marrow. *J. Immunol. Methods* 223, 77–92 (1999).

naturejobs

When I checked into my Bangalore hotel — an unexpectedly lavish establishment — I was politely shown to my room. A few minutes later, a gentleman in a crisp white evening jacket appeared at my door. "Watermelon juice?" he asked. He entered the room and placed it on the desk. "I'm your butler," he said. "If you need me, just call." He pointed to a red button on the wall labelled "Butler". Then he handed me his card. Under his name, in neat script, it read "Butler." I smiled, slightly taken aback. He bowed and grinned as he backed out the door, tray in one hand.

Like much of India, Bangalore is a place of contrasts. My hotel, which caters mostly to business travellers, is part of a service industry stemming from the proliferation of information-technology firms and call centres (and, more recently, biotech companies) in this once-sleepy city. Compared with much of India, Bangalore is wealthy, although the infrastructure doesn't always show it — the population boom has all but overwhelmed the area's traffic-choked roads. In a typically disorganized scenario, the new international airport may open its doors before the road linking it with the city has been completed.

As I sipped my watermelon juice, I considered the prospects for a foreign scientist deciding whether to come to Bangalore, as well as for a native weighing up whether to stay or go abroad. Although still not common, more foreigners have started to work and study at the region's companies and institutions. Going the other way, Uptal Tatu, a biochemist at Bangalore's highly selective Indian Institute of Science told me that representatives of organizations such as Germany's Max Planck Society come to his institute to interview Indians who didn't get accepted, luring some abroad. And Tatu says that none of his PhD students who went abroad for postdocs returned to India.

Still, some Indians are returning to Bangalore (see page 660). And in a few years, the city, building on its biotech growth and pleasant academic campuses, should be an even more appealing destination for foreign scientists — as long as it finishes that road from the airport.

Gene Russo is editor of Naturejobs.

659

Biotech in Bangalore

I t was her beer-brewing know-how that helped make Kiran Mazumdar-Shaw the richest woman in India and sparked a mini biotechnology boom in the country's information-technology capital: her native city, Bangalore.

In 1973, after earning a degree in zoology from Bangalore University, Shaw headed to Australia to study brewing. But work as a brewmaster was hard to find — and being a woman didn't make it any easier. Shaw fortuitously connected with an Irish firm — Biocon Biochemicals — and agreed to start a joint venture in Bangalore in 1978, initially manufacturing enzymes.

"I almost started on the rebound," says Shaw. She was an industry neophyte, but her training in using enzymes to craft the perfect pint provided a robust bridge for her move from beer to biotech. It worked well. In 2007 the company raked in revenues of around US$249 million, a big increase over previous years due in large part to Biocon's sale of its enzyme business to concentrate on pharmaceuticals. Mazumdar-Shaw still has a large stake in Biocon, which is now a publicly traded company.

But even Biocon is still trying to establish itself as a master practitioner of discovery science. At present, most science, engineering and technology jobs in Indian companies emphasize rote tasks rather than encouraging innovation and discovery. Biotech and pharmaceutical companies throughout India are trying to change that (see *Nature* **450,** 580; 2007). The heart of this effort lies in Bangalore, which boasts not only a strength in biotech but a tradition in academic science that stretches back nearly a century. Scientists in industry and academia are trying to recruit and retain top talent, while finding ways to spark innovation.

Arriving in Bangalore, it's not the plethora of IT companies, nor even the sun and the palm trees, that make the greatest impression. It's the traffic. Rickshaws, motorbikes and bicycles weave in and out of the cars

Bangalore has a growing biotech sector, supported by a tradition of academic science. But there are still challenges in committing to discovery science and retaining top talent. **Gene Russo** reports.

Biocon boss Kiran Mazumdar-Shaw has discovery science as her goal.

in what looks like controlled chaos. Trips that should take 20 minutes take more than an hour. Clearly, this is a city whose infrastructure is ill-equipped to handle the tremendous growth of the past decade.

Tortuous paths

A career path in basic research and biotech in India can be equally circuitous. For example, Balasubramanian Gopal, now a young professor at the Indian Institute of Science (IISc) in Bangalore, earned a master's in physics at the Indian Institute of Technology in Kanpur before finding a job in pharma at Torrent Pharmaceuticals in Ahmedabad. He then headed back to school at the IISc for a PhD focusing on structural biology and over to Britain's National Institute of Medical Research for a postdoc in crystallography. Akhilesh Pandey left a sought-after position as a physician in the Indian armed forces to pursue biomedical research in the United States and India (see 'Grand designs'.).

Hosahalli Subramanya, now a vice-president at Bangalore-based Aurigene, a biotech services company, took a risk on Bangalore biotech. After a PhD at the IISc he moved to the University of Oxford, UK, where he completed two postdoctoral stints before returning to India to work for the government-run Central Drug Research Institute at Lucknow. Coming to Aurigene as one its first employees 6 years ago was a gamble. "I had good funding," he says. "And I had offers from other places." Luckily, his move paid off.

Upstart biotechs follow inBangalore's long tradition of academic science, notably the National Centre for Biological Sciences, a small but prestigious institute on the campus of Bangalore's University of Agriculture. And then there's the IISc, founded nearly 100 years ago by entrepreneur Jamsetji Tata, who also started the multinational Tata Group of companies. The

institute has divisions in biological, chemical, physical, mathematical and electrical sciences. And, with increasing government support and its centenary approaching, it plans new buildings for aerospace, biological sciences, information sciences, physical sciences and a centre of excellence in nanoelectronics.

Recruiting talent

The IISc is incredibly selective. Each year about 2,000 students take the IISc entrance test for biological sciences. About 250 are called for interview and 10–12 are admitted as graduate students in each of the five biosciences departments. Even so, the IISc, in common with other Indian institutions, faces a problem: those who demonstrate exceptional talent often leave to pursue a career in the United States or Europe.

Concern about the brain drain prompted the institute to devise an integrated PhD programme that accepts promising undergraduates directly into a PhD, skipping the master's degree, and funds them. IISc chair of biochemistry D. N. Rao admits success has been mixed. Some students have left the PhD programme after the first two years to do a doctorate overseas, taking their newly acquired IISc-sponsored training with them. The shorter graduate training times can also stunt research creativity, says Rao, as investigators work on less risky, more short-term projects because they cannot bank on students devoting many years to the problem. However, the IISc's reputation and Bangalore's increasing global visibility in recent years have started to attract visiting students and new faculty to the campus.

Despite the tough competition, graduate student recruits often lack important communications skills, says Gopal. Many come originally from villages and have poor English, the accepted language of science in India. "They have an inferiority complex, and keep quiet," he says. "But when you actually look at them at the bench, you say 'wow'." Gopal calls them the "H1B missed" generation, referring to the visa required to work in the United States. They have neither the language skills nor the funding to go abroad.

Well-qualified people are needed by a biotech sector that is still recruiting. Subramanya says Aurigene plans to take on 60–70 employees in the next few months on top of the 250 scientists it already has. He says some specialist positions will be difficult to fill. The bustling IT sector also competes for talent, especially for entry-

level positions with a computer-skills component. The strategy of contract-research giant Quintiles Transnational, in Bangalore since 1997, is to recruit college graduates in the pharmaceutical and life sciences and train them to process clinical research data and arrange clinical trials for big pharma clients.

Taking a risk

But the biggest challenge facing Bangalore biotechs is to establish a culture of innovation and invest in higher-risk discovery research. Subramanya believes that the stigma that India can do only copycat research is no longer there. Six years ago, Aurigene would be expected to just make a given compound, he says. Now the company has more of an R&D focus — trying to attack a target with a compound that has the right properties to be effective. Laws passed in 2005 have made it easier to protect intellectual property in India, a crucial step to curing the private sector of risk aversion.

At Biocon, Mazumdar-Shaw says discovery science is a major goal, as is collaboration with the area's much smaller biotechs. She claims that the company has fared so well because of its capacity to expand and evolve its repertoire — from enzymes to biopharmaceuticals and monoclonal antibodies. Mazumdar-Shaw, a member of a special panel of the Indian government's Department of Biotechnology working to reform bioscience in the country, says schooling will be key. Specialist training before going into industry and management training as part of biotechnology education are both in the works. "As a country of 1.2 billion, we still haven't made our own drug," says Gopal, downplaying generics and herbals. "That still drives me."

Gene Russo is editor of NatureJobs.

GRAND DESIGNS

In 1990, Akhilesh Pandey was a promising physician, having graduated from India's prestigious Armed Forces Medical College in Pune. But he convinced his parents to help pay the fee releasing him from his 20-year bond of service in the Indian armed forces. They had to sell their car to do so. Pandey had decided that basic research was his real passion, and his parents trusted his instincts. Twelve years later he would gamble again, in the hope of establishing an independent non-profit scientific institute in India, a rarity in the country.

After securing his exit, Pandey headed to the University of Michigan to get his PhD and then finished his residency at Brigham and Women's Hospital in Boston, Massachusetts. He followed that with a postdoc at the Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, and a visiting-scientist position at the University of Southern Denmark in Odense. It was during his time at Johns Hopkins School of Medicine in Baltimore, Maryland, where Pandey is still an associate professor, that he devised his ambitious plan for a bioinformatics institute.

Pandey envisaged an Institute of Bioinformatics and wanted it to be in Bangalore. The primary project would be a huge human protein database, with information on protein interactions, modifications and roles in disease. Again his family came to his aid. Pandey got loans from his brother and parents, and maxed out his credit cards. His Human Protein Reference Database was sustained by a grant from the US National Center for Biotechnology Information, under the auspices of the National Institutes of Health (NIH) in Bethesda, Maryland.

Six years later he has been able to pay back his family, but not himself — and there are no guarantees of the institute's future. Expansion is difficult because lenders in India are wary of an institute lacking steady government or venture-capital backing. Pandey had to convince his own lawyer that the institute wasn't a way to surreptitiously launder money.

At its peak, the institute had 70 employees; it now has 50 since losing funding as a result of the NIH's budget cuts. Nevertheless, Pandey, who splits his time between Johns Hopkins and the institute, plans to construct a mass spectrometry facility as well as a community-driven proteomics initiative called Human Proteopedia. "We want to spoil the user," he says. Now that the institute has proved its bioinformatics mettle, he and his colleagues are seeking new grants. He is optimistic that the institute can flourish — and that he'll be able to pay off the last of his debt. **G.R**.

> **"As a country of 1.2 billion, we still haven't made our own drug. That still drives me."** — Balasubramanian Gopal

New beginnings: staff of the Institute of Bioinformatics, Bangalore.

A. PANDEY

MOVERS

Dave Tapolczay, chief executive, MRC Technology, London

2007-08 Chief scientific officer, SAFC Pharma, Cambridge, UK
2003-07 Vice-president, GlaxoSmithKline Pharmaceuticals, Harlow, UK
2000-03 Vice-president, Millennium Pharmaceuticals, Boston, Massachusetts

Dave Tapolczay knows a thing or two about new technologies. Early in his career, he developed chemical compounds for use in insecticides and, later, for pharmaceutical targets. More recently, he's helped to commercialize technologies that automate chemical processes. He has launched three biotechnology companies, a consulting firm and an investment company. And now he is taking his skill at spotting promising technology to the UK government sector as the new chief executive officer of MRC Technology (MRCT), where he'll take charge of technology transfer.

In 1997, Allan Marchington, a venture capitalist with Apposite Capital in London, discussed a start-up venture with Tapolczay. Marchington says that within two minutes he knew he'd found the right person. Tapolczay had an imaginative approach to the nascent company's portfolio. "He could see the best opportunity to make commercial sense from the science," said Marchington. In 2000, they sold the company to Millennium Pharmaceuticals, and Tapolczay moved to Boston, Massachusetts, to take responsibility for the development of the entire Millennium pipeline. During his tenure, Millennium introduced Velcade, the first proteasome inhibitor marketed, and a second- or third-line treatment for multiple myeloma.

"I've had a quite varied career," says Tapolczay, noting that each sector has its appeal. In large corporations, it is the opportunity to work on cutting-edge, innovative science. For biotechnology start-ups, it is the "adrenaline of trying to keep the company alive and make it a success". Tapolczay earned a BSc and a PhD in chemistry from the University of Southampton, UK. After a postdoctoral fellowship at the University of Oxford, he got his start in industry at G. D. Searle Pharmaceuticals, and then moved to ICI.

Tapolczay has several goals in his new post: raising royalty revenue (MRCT had income of £46 million (US$92 million) in fiscal year 2006–07, all of which goes back into MRC research); spinning off companies from the MRC (there have been 17 in the past 20 years); licensing products to existing companies; and using new disease biomarkers, diagnostics and therapies to improve health.

Tapolczay is also keen to explore new licensing opportunities with US and Japanese companies. Still, he says the majority of his effort will be dedicated to working directly with MRC-funded institutions. "One thing I'll do is undertake an audit," says Tapolczay. He envisages rating performance on a scale of 1 to 10. "If it's a 5," he says, "I want to know how we can do better."
Jill U. Adams

NETWORKS & SUPPORT

Göttingen bridges the gap

A new doctoral training programme will tackle seriously complex problems. Called the 'Physics of biological and complex systems', the programme is part of the International Max Planck Research Schools (IMPRS), a group of interdisciplinary centres of excellence run by the Max Planck Institutes in conjunction with German universities.

The new programme will build on the University of Göttingen's expertise in physics and will focus on, for example, the timing of neural networks and the physics of cardiac dynamics. It is a joint venture between two Göttingen-based Max Planck Institutes and the university. The Max Planck Society is devoting approximately €4 million (US$6.2 million) to the programme over the next 6 years.

Helmut Grubmüller, a director at the Max Planck Institute for Biophysical Chemistry, oversees the programme and extols its breadth and structure. It will focus both on using basic physics to solve existing biological problems and researching the unique physics of biological systems. Other programmes tend to keep the topics separate. This structured, fast-track PhD programme will involve special courses, lab rotations and methods tutorials to help students earn their PhDs within 3.5 years compared to 4 or more.

Recent recruitments, including Grubmüller and Eberhard Bodenschatz, a director at the Max Planck Institute for Dynamics and Self-Organization, have strengthened the technical expertise in super-resolution optical microscopy. For example, students will learn to image intracellular structures to a resolution of 15 nanometres and perform single-molecule spectroscopy.

Such expertise will help the new programme address the growing interest in nanometre-scale biological techniques. Bodenschatz believes this innovative training will form a bridge between Göttingen's strong physics and biosciences communities. Although the official start date for recruiting students to the IMPRS is 1 May, five of the 30 slots are already filled.

The IMPRS programmes are part of Germany's ongoing efforts to strengthen graduate training. Of the 49 existing schemes, two — the IMPRS for molecular biology and for neurosciences — have already proved successful at Göttingen. The molecular biology programme was recognized in 2006 as one of the top 10 international master's courses in Germany by the German Academic Exchange Service.
Virginia Gewin

POSTDOC JOURNAL

Getting that lucky break

"If you can't get a project to work," a fellow postdoc stated, "you're no good." Her comment, expressed during an informal discussion with senior management, was undisputed. I pondered how many failed projects are permitted before a budding scientist is dismissed as "no good".

During the past two years, three of my projects have ended prematurely thanks to negative results and reagent problems. My objective now is to discover the role of a new protein. But more than a year later, it seems that my original hypothesis is wrong, and I am left questioning my choice of projects.
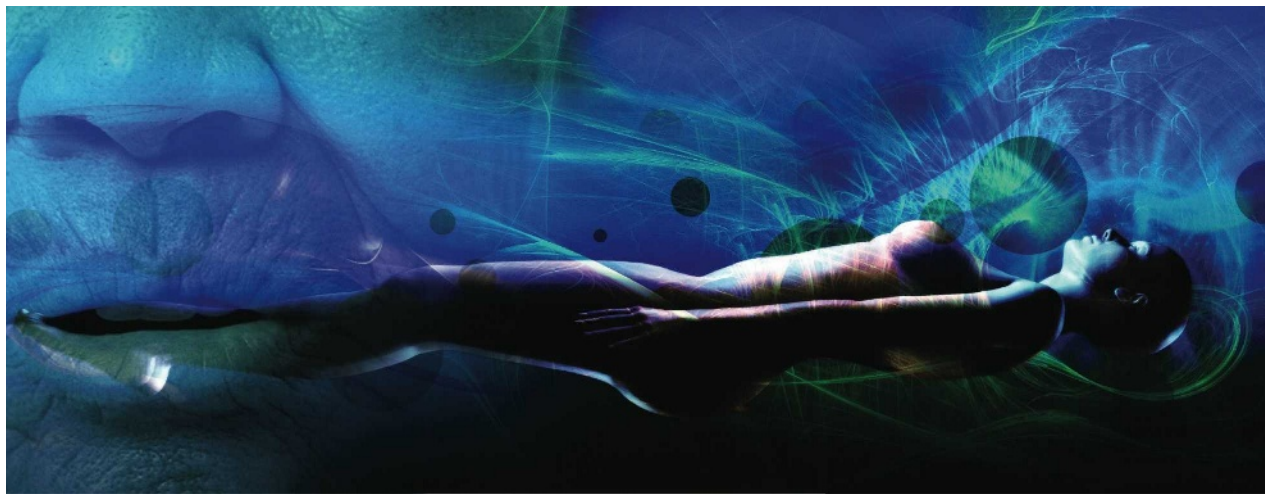
On Singapore's Biopolis campus where I work, fields with great potential for therapy and profit are strongly encouraged. A senior scientist's sole question about my work underscored this emphasis: "Is it big?" he asked, alluding to its chances of being published in a high-impact-factor journal. Considering that the direction of my project remains unclear, let alone its long-term prospects, well-meaning investigators have questioned whether I should stick with it. Indeed, how do I know when to cut my losses?

One senior scientist told me that a successful project comes from "sheer luck". I'm not usually superstitious, but I welcomed this Chinese New Year by switching on all the lights in my apartment to summon the God of Fortune.
Amanda Goh is a postdoctoral fellow in cell biology under the Agency of Science, Technology and Research in Singapore.

Shoppers

No purchase necessary.

James Patrick Kelly

They pass down the row of shops, not exactly together. Rick keeps wandering ahead of her and then pauses to ogle the display windows. He does not notice the cold. He will never notice the cold. Rick lingers over an array of designer sheets in a linen store. He's dressed for another place and a different time, black crushed-velvet jacket, glitter shirt with a whisper of lace at the sleeves, gabardine trousers. He has a smooth young man's face with skin of fourteen carat gold.

She drags herself after him. She is stooped and as grey as the dirty snow caked along the gutter. When she wheezes, she can see the breath curling out of her mouth into the chill air. Her bulky fur coat clamps her in a bear hug. She slides to his side, feeling for black ice with tiptoes as if it might bite her. If she falls, she might never get up.

"That one." Rick points at a set of Bellino sheets the colour of blood oranges. "Buy me that." A label opens on the shop's window: sateen weave thousand thread count Egyptian cotton.

"Please." She leans into him as if to propel him down the street. "You have enough."

"You could be sleeping on them." He brushes a hand against her withered cheek. "You."

They stop and start their strolling tango past a dozen stores. She waits outside while he tries on some artful scars at Malade and buys a mink hat at Skarsgård's that costs as much as a geriatric nurse makes in a year. He teases her in front of an ice-cream shop.

"Remember pistachio?" says Rick.

"Remember when snacks weren't an IV drip?"

"Why are you being mean?" Her disposable eyes get bright and a little misty; the latest Cibas come with a special-effects package. "It's my fault I'm a hundred and eighty-three?"

"You chose." Rick slips an arm around her waist. "You're still choosing."

Their appointment at Evergreens isn't until three and they are early. Rick takes her coat and hangs it up and she sags onto a chair in the waiting room. He leaves her so he can poke around, picking up fingers from their display cases, aiming them at her like guns. There are medfingers, chatfingers, dofingers, sexfingers, mindfingers and glowfingers. There are fingers of rain forest mahogany and fingers of carbon allotropes stronger than high-tensile steel. He reads the label on the vintage Bösendorfer handset in a display case. Its fingers can play Ravel's *Piano Concerto for the Left Hand* or the single-hand études of Saint-Saëns.

A salesgirl in a white smock appears at the door at the far end of the room. "Master Evergreen will see you now." She carries herself with the unconscious ease of someone still in her birth body.

Rick pulls the sexfinger from his manual unit and replaces it in the display case. He crosses the showroom to her.

She tries to get up but she can't. Rick and the salesgirl each take an arm and guide her into the showroom.

Evergreen is a vision of white teeth, dark polyskin and post-retro body design. It is happy to see them, or at least as happy as a bot can get.

"So we're back then," it says. "Are we any closer to a decision?"

She brushes past it to the two naked bodies suspended from the neural web. One is sleek and hard and eternal; the other soft and warm and human. She touches the chin of the flesh body.

"As I said last time, these are just base models." Evergreen comes up behind her. "We have many, many options to consider before you make the transfer."

"Will it hurt?"

Evergreen shrugs. "A passing moment in a long, long life."

She snorts in disbelief and turns to him. Rick is staring at the glowing breasts of the bot body, his face flushed bronze. There is a word on his lips that she can't quite make out.

"What do you think?" she says.

Rick starts as if roused from a daydream. "Don't ask me." He sticks his hand into the pocket of his jacket. "I'd love either one." He turns the pocket inside out and flicks lint onto the showroom floor.

Evergreen's smile flickers as if from a power surge, and then brightens to maximum again. "Shall we step into my office then?"

She shivers. "It's a big decision," she says. "I guess I'm still looking." Then she tugs at the sleeve of Rick's jacket.

He scowls. "This can't go on."

"But it can." She pinches his golden cheek.

Reluctantly, he offers her his arm. "You're wasting our time."

She propels him toward the exit. "But it's mine to waste, dear." ∎

James Patrick Kelly has won the Hugo, Nebula and Locus Awards. He podcasts free readings of his work at http://feeds. feedburner.com/freereads. *The Wreck of the Godspeed*, **a new collection, is forthcoming in August.**